

JALA TECH:

DATA SCIENCE CHALLENGE

Created by:

Muhamad Rifki Taufik




PREFACE AND OUTLINE

Report ini disusun untuk menjawab tantangan data science yang dilakukan oleh Jala Tech. Adapaun dataset yang diberikan pada applicant sebanyak 3 macam yaitu, siklus, kualitas air, dan budidaya. Seluruh analisis dan visualisasi menggunakan program R dengan package yang dipergunakan yaitu tidyverse, ggplot2, naniar, dan fmsb. Berikut merupakan outline laporan yang disusun penulis:

1. Siklus tiap kolam
 - a. Deteksi anomaly, kalkulasi luas dan volume kolam, identifikasi lamanya siklus
 - b. Analisis (lama siklus, luas kolam, dan jumlah benur tiap daerah)
2. Kualitas air
 - a. Identifikasi missing value
 - b. Suhu air
 - c. Oksigen terlarut
 - d. Salinitas
 - e. pH
3. Budidaya udang
 - a. Identifikasi missing value
 - b. Productivity
 - c. Survival rate
 - d. Feed Conversion Rate (FCR)
4. Kesimpulan

Disclaimer: Penulis tidak memiliki background budidaya udang, sehingga interpretasi data merupakan hasil persepsi dari penulis.



1. Siklus tiap kolam

Dalam dataset daftar_siklus.csv terdapat 5 baris dengan 9 colom/variable yang terdiri dari Kode Siklus, Kode Kolam, Panjang Kolam (m), Lebar Kolam (m), Kedalaman Kolam (m), Tanggal Tebar, Tanggal Selesai Siklus, Jumlah Benur (ekor), dan Daerah.

a. Deteksi anomaly, kalkulasi luas dan volume kolam, identifikasi lamanya siklus

Kode.Siklus	Kode.Kolam	Panjang.Kolam	Lebar.Kolam	Kedalaman.Kolam
Min. : 50.0	Min. : 3	Min. : 36.20	Min. : 36.30	Min. : 0.9
1st Qu.: 66.0	1st Qu.: 4	1st Qu.: 49.90	1st Qu.: 46.40	1st Qu.: 1.1
Median : 295.0	Median : 2266	Median : 79.16	Median : 59.37	Median : 1.5
Mean : 304.2	Mean : 1822	Mean : 71.98	Mean : 57.61	Mean : 21.0
3rd Qu.: 298.0	3rd Qu.: 3418	3rd Qu.: 79.16	3rd Qu.: 59.37	3rd Qu.: 1.5
Max. : 812.0	Max. : 3421	Max. : 115.47	Max. : 86.60	Max. : 100.0
Tanggal.Tebar	Tanggal.Selesai.Siklus	Jumlah.Benur	Daerah	
Length:5	Length:5	Min. : 100000	Length:5	
Class :character	Class :character	1st Qu.: 175000	Class :character	
Mode :character	Mode :character	Median : 200000	Mode :character	
		Mean : 329434		
		3rd Qu.: 583740		
		Max. : 588432		

Figure 1 analisis deskriptif dataset siklus

Pada Figure 1, ditemukan sebuah anomaly pada variable kedalaman kolam, dimana menunjukkan kedalaman 100, dimana yang lain hanya sekitar 1 meter, sehingga diasumsikan ini adalah salah input/human error, maka data 100 ini diubah menjadi 100 cm = 1 meter.

Kemudian, karena dirasa perlu maka dibuat 3 variabel baru, yaitu luas (perkalian panjang dan lebar kolam), volume (perkalian luas dan kedalaman kolam), dan lamanya siklus (tanggal selesai siklus dikurangi tanggal tebar, dalam satuan hari).

vol	luas	lamaSiklus
Min. : 1630	Min. : 1680	Min. : 57.0
1st Qu.: 1848	1st Qu.: 1811	1st Qu.: 59.0
Median : 7050	Median : 4700	Median : 86.0
Mean : 203510	Mean : 4578	Mean : 86.6
3rd Qu.: 7050	3rd Qu.: 4700	3rd Qu.: 111.0
Max. : 999970	Max. : 10000	Max. : 120.0

Figure 2 analisis deskriptif variabel baru: luas, volume, lama siklus

b. Analisis (lama siklus, luas kolam, dan jumlah benur tiap daerah)

Figure 3 menunjukkan karakteristik siklus di tiap daerah. Terdapat hal yang menarik disini, Nampak penyebaran jumlah benur di tiap kolam tidak bergantung dengan luasnya kolam. Hal ini ditunjukkan ketika luas kolam di Bantul mencapai 10000 m², jumlah benur yang disebar hanya sekitar 100.000 ekor namun pada kolam di Lampung Selatan dengan luas hanya 4699 m², benur yang disebar sebanyak 583740 ekor. Hal ini mengindikasikan bahwa para pengelola tambak belum memperhatikan betul proporsi jumlah benur yang disebar berdasarkan luas dari masing-masing kolam.

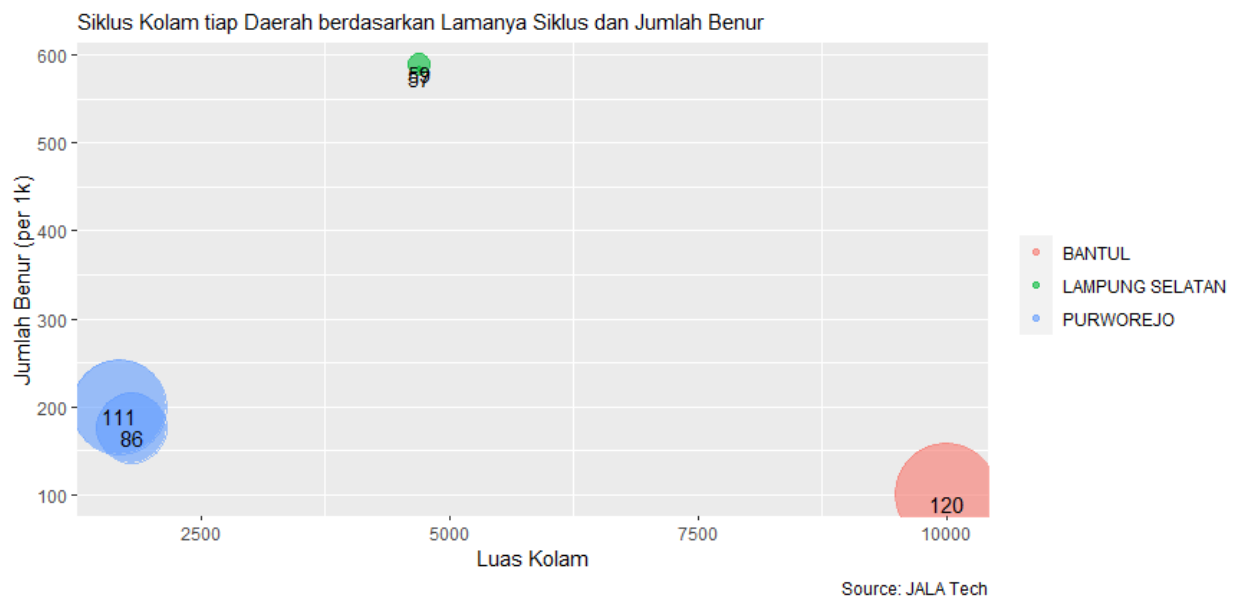


Figure 3 analisis karakteristik siklus

Disisi lain, Nampak bahwa lamanya siklus juga sangat bervariasi, tidak bergantung dengan jumlah benur dan luasnya kolam. Pada kolam di Lampung Selatan, walau jumlah benur yang -disebar paling banyak, namun lamanya siklus berkisar 57-59 hari. Sedangkan untuk benur yang paling sedikit justru malah memerlukan waktu yang paling lama yaitu 120 hari. Hal ini menunjukkan bahwa perlu identifikasi lebih jauh penentuan jumlah benur dengan mempertimbangkan luas kolam dan perkiraan lamanya siklus untuk mendapatkan hasil yang optimal.

Daerah	Jumlah Benur	Luas	Lama Siklus
Bantul	100000	9999.70	120
Purworejo	200000	1679.68	111
Lampung Selatan	583740	4699.73	57
Lampung Selatan	588432	4699.73	59
Purworejo	175000	1811.37	86

2. Kualitas air

Dataset terdiri dari 581 record dengan 6 variable, yaitu Waktu Pengukuran, Kode Siklus, Suhu Air, Oksigen Terlarut, Salinitas, pH. Kemudian dengan bantuan dataset awal, dibuat suatu variable baru yaitu kolam, yang berisikan daerah dan no kolam (Bantul, LamSel 1, LamSel2, Purworejo 1 dan Purworejo 2) dg tujuan untuk mempermudah analisis. Kemudian Kode Siklus dihilangkan karena sudah digantikan dengan variable Kolam.

Waktu. Pengukuran	Suhu. Air	Oksigen. Terlarut
Min. : 2018-07-16	Min. : 24.00	Min. : 0.00
1st Qu.: 2018-08-23	1st Qu.: 27.00	1st Qu.: 4.45
Median : 2018-11-06	Median : 28.31	Median : 5.34
Mean : 2018-12-25	Mean : 28.67	Mean : 5.63
3rd Qu.: 2019-03-05	3rd Qu.: 29.62	3rd Qu.: 6.40
Max. : 2019-07-31	Max. : 127.00	Max. : 18.26
	NA's : 3	NA's : 12
Salinitas	pH	kolam
Min. : 0.00	Min. : 2.750	Bantul : 136
1st Qu.: 27.38	1st Qu.: 7.530	LamSel 1 : 70
Median : 30.91	Median : 7.860	LamSel 2 : 84
Mean : 31.00	Mean : 7.704	Purworejo 1: 157
3rd Qu.: 37.16	3rd Qu.: 8.240	Purworejo 2: 134
Max. : 39.74	Max. : 9.770	
NA's : 60	NA's : 5	

Figure 4 analisis deskriptif kualitas air

Terlihat secara analisis awal terdapat ada beberapa NA data, maka perlu diidentifikasi lebih jauh terkait missing value ini sebelum dilakukannya analisis.

a. Identifikasi missing value

Missing value merupakan tidak adanya data/nilai dalam record table yang telah tersusun. Hal ini dapat menjadi evaluasi mengapa hal ini terjadi, atau jika missing value masih dalam batas tolerance maka missing value bisa dihandle atau diomit agar ketika dilakukan analisis, hasil tidak memberikan arah kesimpulan yang salah (misleading decision).

Missing Value Record		
Indicator	#	%
Terdapat missing value	TRUE	
Berapa banyak	80	2.29%
Waktu Pengukuran	0	0%
Kolam	0	0%
Suhu Air	3	0.51%
Oksigen Terlarut	12	2.07%
Salinitas	60	10.30%
pH	5	0.86%

Nampak pada table diatas bahwa terdapat missing value (NA) sebesar 2.29% dari total record yang ada. Figure 5 menunjukkan bahwa salinitas memiliki missing value terbanyak yaitu sebesar 10.3%. hal ini bisa menjadi evaluasi manajemen, mengapa record dari salinitas bisa banyak yang

hilang/ tidak tersedia.

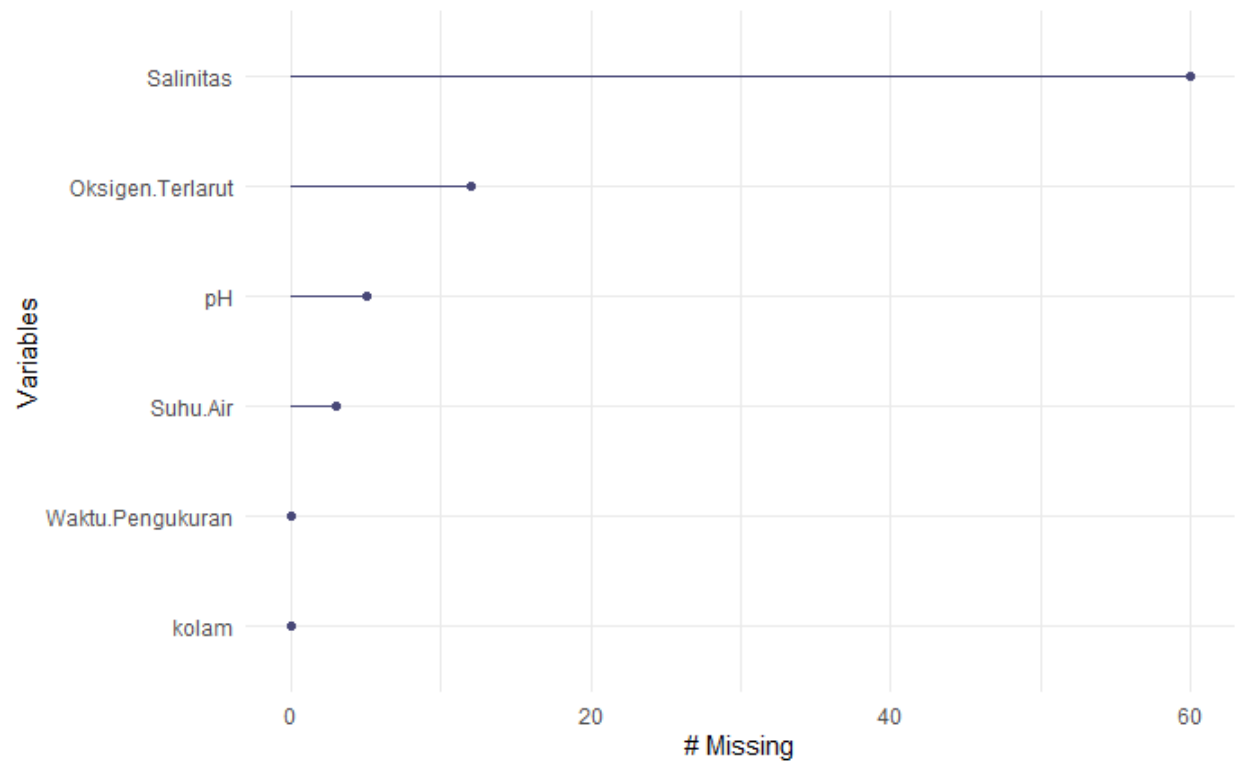


Figure 5 numbers of missing value in each variable

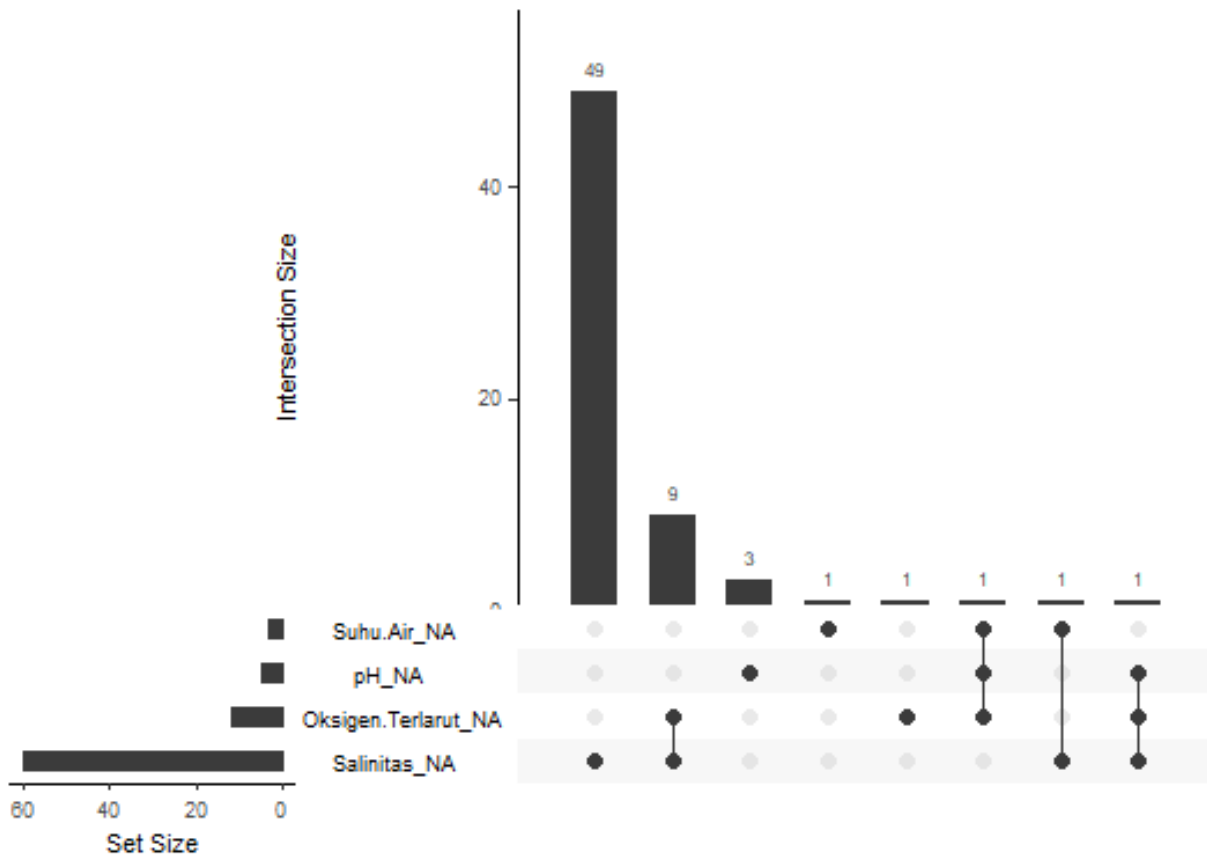


Figure 6 intersection missing value: persebaran NA yang terjadi

Sebaran dari missing value teridentifikasi bahwa hanya variable salinitas yang sering terjadi, sehingga tidak ada kaitan dengan variable lain yang mungkin mempengaruhi terjadinya missing value pada salinitas. Pada Figure 7, terlihat bahwa kolam yang memiliki record salinitas yang

hilang/tidak tersedia yaitu pada kolam Purworejo 2, yang kemudian perlu adanya peninjauan kembali mengapa missing value banyak terjadi hanya pada kolam ini.

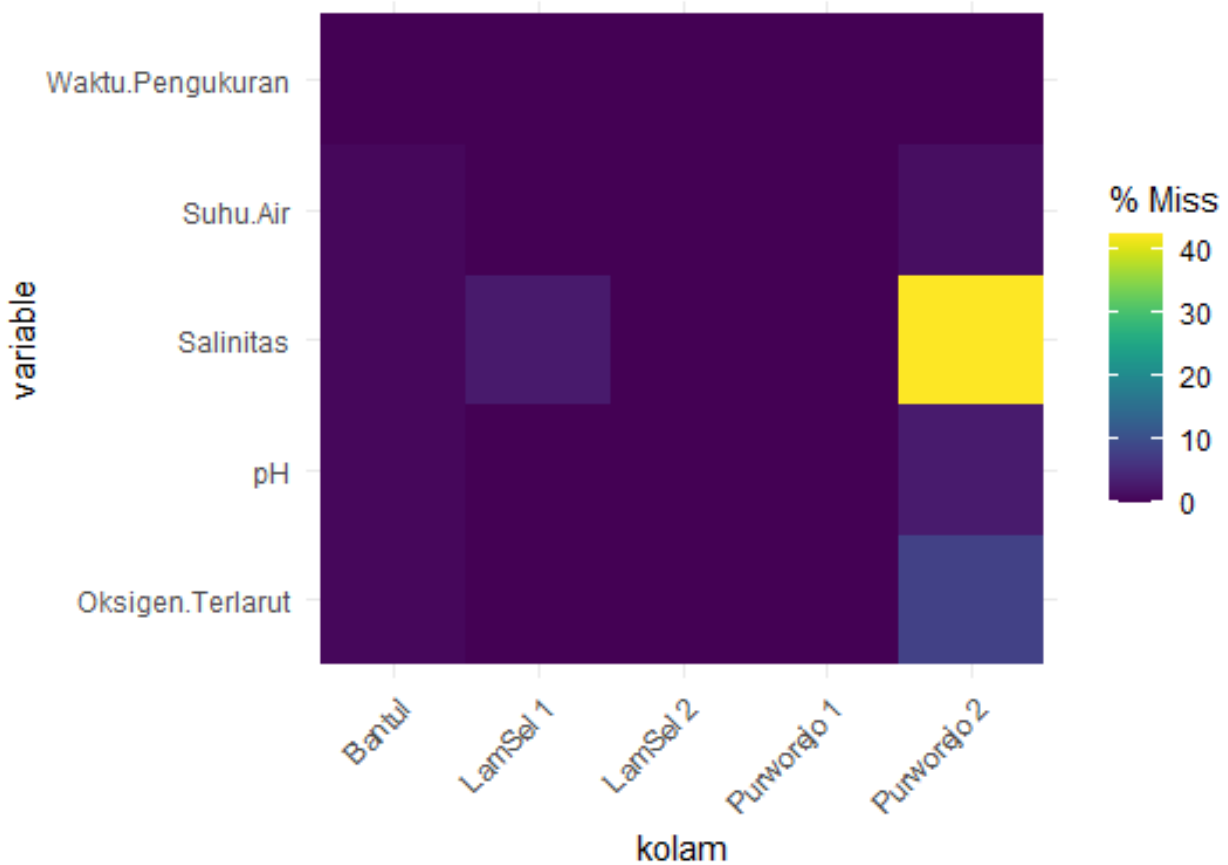


Figure 7 heatmap sebaran missing value

b. Suhu air

Suhu air tersebar seperti pada Figure 8 dan teridentifikasi adanya extreme value yang kemungkinan adalah outlier. Outlier ini merupakan data anomaly yang mungkin terjadi akan 2

hal, yaitu adanya extreme value (nilai extreme) yang dikarenakan adanya extreme condition atau dikarenakan alat ukur yang error. Untuk kepentingan analisis maka outlier ini dihilangkan (omit).

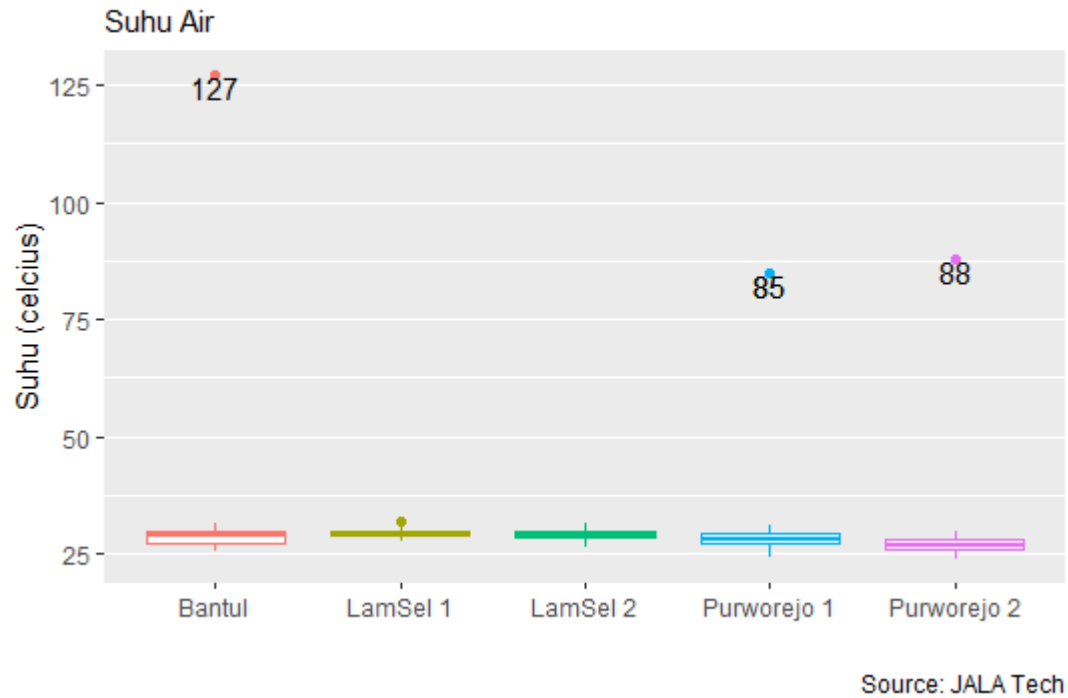
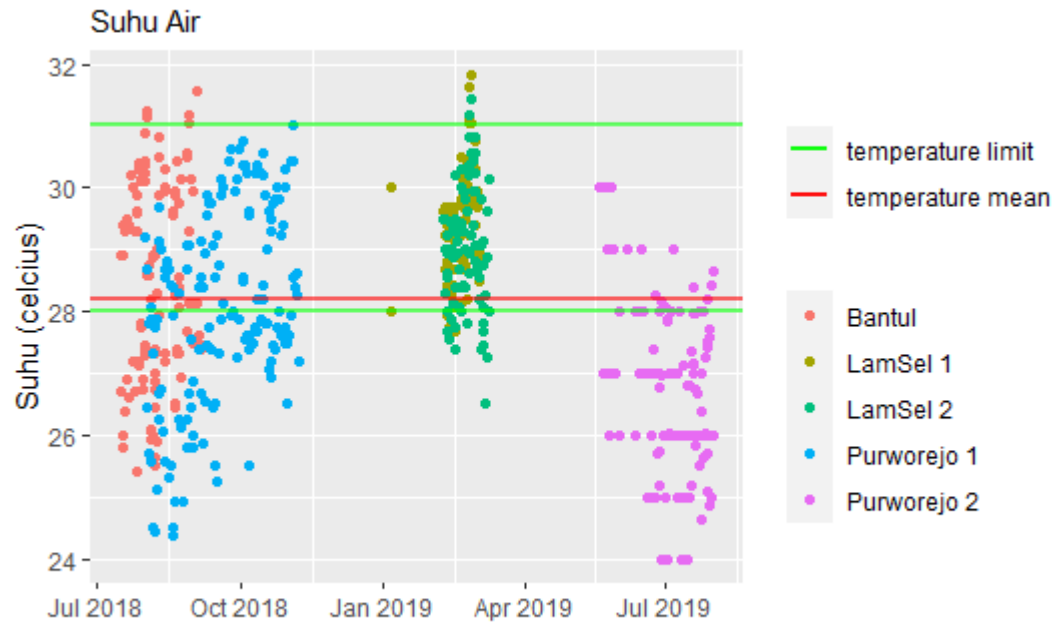


Figure 8 identifikasi outlier dalam data suhu air

Figure 9 menunjukkan distribusi suhu air setelah outlier dihilangkan dimana terdapat batas bawah dan batas atas suhu air yang telah ditentukan sebelumnya pada range 28-31 derajat

celcius. Nampak bahwa masih sangat banyak record yang tidak dalam range yang telah ditentukan.



Source: JALA Tech

Figure 9 persebaran data suhu air pada batasan normal

Masing-masing kolam pasti pernah diluar suhu yang sudah ditentukan. Namun teridentifikasi bahwa ada 3 dari 5 kolam lebih dari 40% record. Hal ini menjadi perhatian bahwa pengelolaan kolam banyak kondisi diluar suhu air yang dianjurkan.

Kolam	#	%
Bantul	59	44.02
Lampung Selatan 1	9	12.86
Lampung Selatan 2	14	16.67
Purworejo 1	79	50.97
Purworejo 2	97	74.04

c. Oksigen terlarut

Pada sebaran oksigen terlarut tidak ditemukannya extreme value yang sangat berbeda seperti yang terjadi pada suhu air. Meski demikian banyak ditemukannya outlier di kolam bantul dan purworejo 1 seperti yang dilihatkan pada Figure 10.

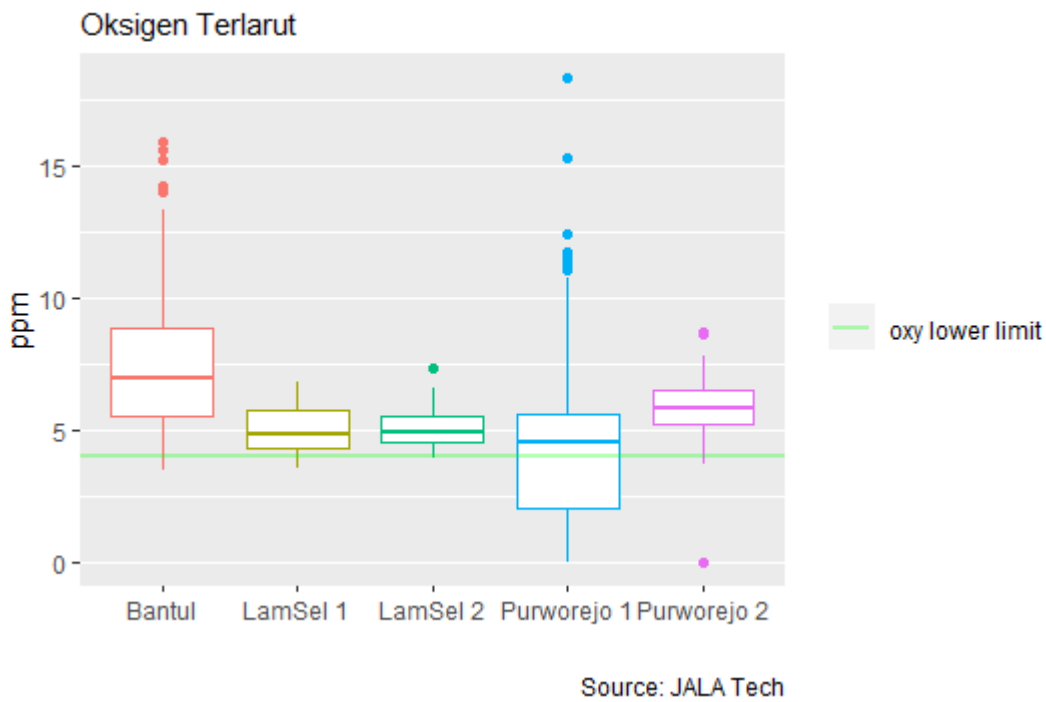


Figure 10 distribusi oksigen terlarut

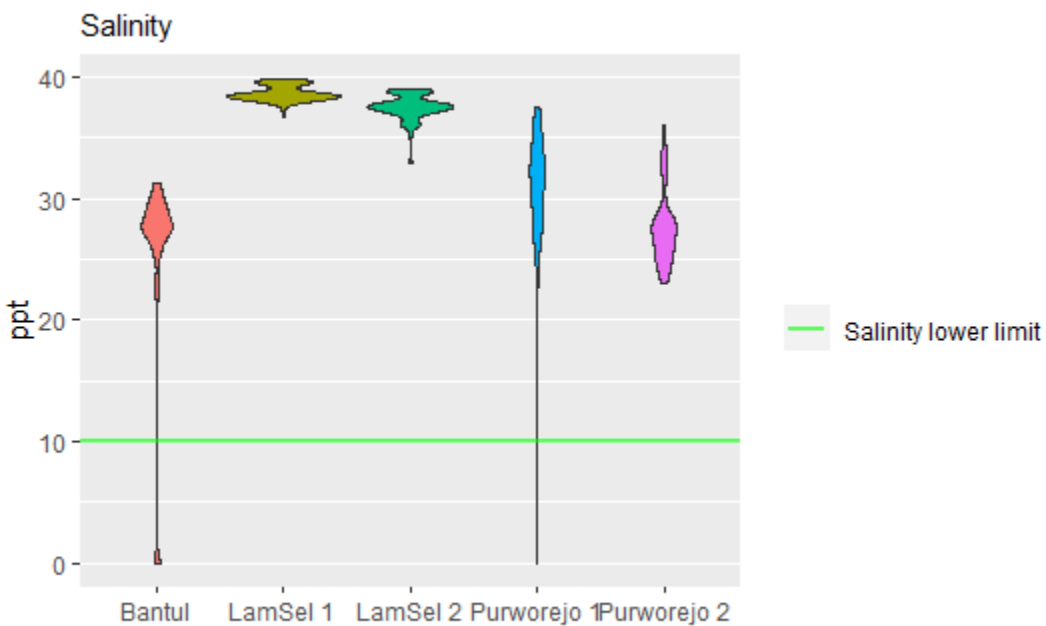
Banyaknya record yang kurang dari kadar oksigen terlarut kurang dari 4 ppm ditunjukkan pada table dibawah berikut. Kolam Purworejo 1 merupakan kolam yang record oksigen terlarutnya paling diluar batas oksigen terlarut yang dibutuhkan udang, yaitu sebesar 40%.

Kolam	#	%
Bantul	7	5.18
Lampung Selatan 1	9	12.85

Lampung Selatan 2	1	2.38
Purworejo 1	63	40.12
Purworejo 2	3	2.44

d. Salinitas

Batas salinitas yang ditentukan sebesar 10 ppt. namun untuk dua kolam di Bantul dan Purworejo 1, terdapat beberapa record yang kurang dari batas minimal yang ditentukan.



Source: JALA Tech

Figure 11 distribusi salinitas

Kolam	#	%
Bantul	11	8.14
Lampung Selatan 1	0	0
Lampung Selatan 2	0	0

Purworejo 1	3	1.91
Purworejo 2	0	0

e. pH

Batas minimal pH air yang ditentukan yaitu kisaran 7.5-8.5 dengan density 5. Adapun sebaran pH masing-masing kolam ditunjukkan pada Figure dan Tabel dibawah ini. Keduanya menunjukkan bahwa setiap kolam memiliki tingkat pH yang sudah ditentukan secara konsisten. Meski terjadi perubahan dibawah dan diluar batas yang ditentukan, namun density yang ada masih dibatas toleransi yaitu kurang dari 5.

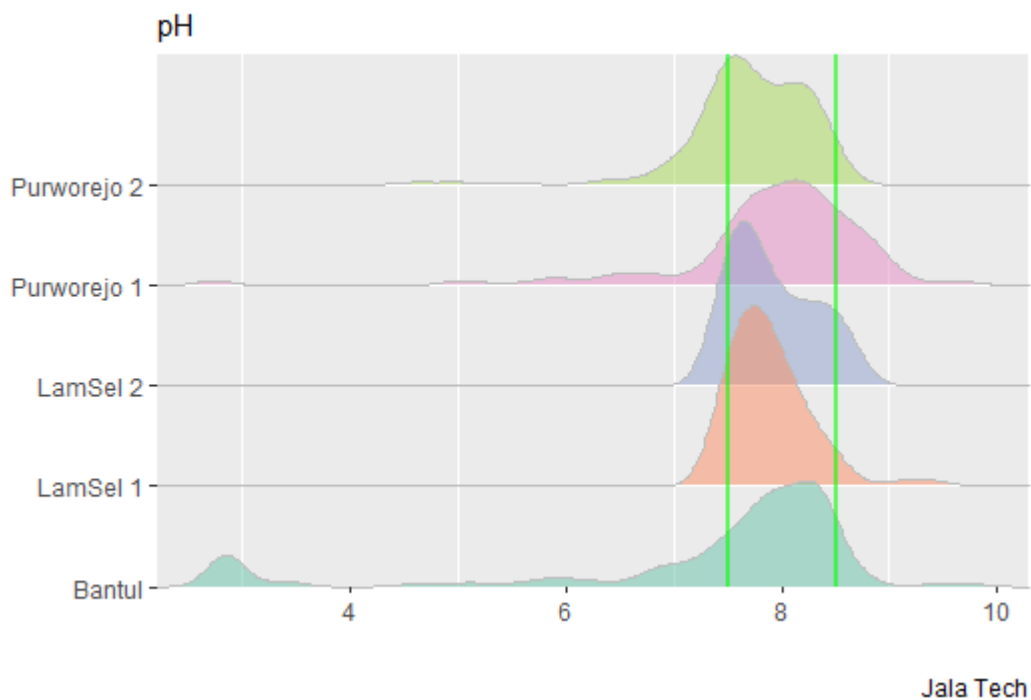


Figure 12 distribusi pH

Kolam	avg	N	Sd	Se
Bantul	7.27	135	1.67	0.143

Lampung Selatan 1	7.89	70	0.37	0.044
Lampung Selatan 2	7.92	84	0.37	0.041
Purworejo 1	7.88	157	0.96	0.077
Purworejo 2	7.70	130	0.60	0.053

3. Budidaya udang

Total observasi yang ada dalam dataset ini sebanyak 438 dengan 11 variables, yang terdiri dari Kode Siklus, Age, Date, ABW, Size, Total Weight, Feed, Feed Accumulation, Survival Rate, Panen, Berat Panen. Pada dataset ini pun dilakukan sama dengan dataset kualitas air, dimana dimerge variable nama kolam berdasarkan kode siklus yang ada.

Age	Date	ABW	Size
Min. : 0.00	Length:438	Min. : 2.050	Min. : 50.00
1st Qu.: 21.25	Class :character	1st Qu.: 5.185	1st Qu.: 77.25
Median : 43.00	Mode :character	Median : 7.165	Median :139.50
Mean : 47.12		Mean : 8.921	Mean :163.85
3rd Qu.: 69.75		3rd Qu.:12.988	3rd Qu.:192.75
Max. :120.00		Max. :20.000	Max. :488.00
		NA's :404	NA's :404
Total.Weight	Feed	Feed.Accumulation	Survival.Rate
Min. : 0.0	Min. : 2.00	Min. : 0	Min. : 28.77
1st Qu.: 383.1	1st Qu.: 17.00	1st Qu.: 264	1st Qu.: 77.94
Median : 858.3	Median : 25.00	Median : 559	Median : 81.83
Mean :1217.3	Mean : 49.22	Mean :1227	Mean : 83.14
3rd Qu.:1647.2	3rd Qu.: 59.00	3rd Qu.:1640	3rd Qu.: 94.53
Max. :4749.1	Max. :200.00	Max. :6021	Max. :140.30
NA's :404	NA's :78		
Panen	Berat.Panen	luas	Suhu.Air
Min. : 28861	Min. : 379.6	Min. : 1680	Min. :27.22
1st Qu.: 43471	1st Qu.: 479.3	1st Qu.: 1680	1st Qu.:28.81
Median : 50353	Median : 532.8	Median : 4700	Median :29.02
Mean :122945	Mean :1213.1	Mean : 4818	Mean :28.74
3rd Qu.: 76046	3rd Qu.: 822.3	3rd Qu.:10000	3rd Qu.:29.30
Max. :542368	Max. :4975.9	Max. :10000	Max. :29.43
NA's :431	NA's :431		
Oksigen.Terlarut	Salinitas	pH	kolam
Min. :4.597	Min. :25.38	Min. :7.271	Bantul :121
1st Qu.:4.597	1st Qu.:25.38	1st Qu.:7.271	LamSel 1 : 58
Median :5.040	Median :30.71	Median :7.883	LamSel 2 : 60
Mean :5.717	Mean :30.60	Mean :7.683	Purworejo 1:112
3rd Qu.:7.387	3rd Qu.:37.53	3rd Qu.:7.891	Purworejo 2: 87
Max. :7.387	Max. :38.64	Max. :7.920	

Figure 13 analisis deskriptif dataset budidaya

Data ini terdapat anomaly yang sangat Nampak, dikarena teridentifikasi missing value yang sangat besar di beberapa variable.

a. Identifikasi missing value

Sangat disayangkan bahwa beberapa variable memiliki tingkat missing value yang sangat tinggi dimana record yang hilang mencapai 90%. Hal ini menjadi tanda tanya besar, apakah ada kesalahan pengukuran, atau memang sudah sesuai dengan prosedur? Nampaknya, ada beberapa variable yang memang hanya diukur saat panen saja, hal ini menjadi rekomendasi untuk data administrator untuk dapat membuat satu dataset terpisah untuk record panen tiap kolam.

Missing Value Record		
Indicator	#	%
Terdapat missing value	TRUE	
Berapa banyak	2152	30.7%
Age	0	0%
Date	0	0%
ABW	404	92.2%
Size	404	92.2%
Total Weight	404	92.2%
Feed	78	17.8%
Feed Accumulation	0	0%
Survival Rate	0	0%
Panen	431	98.4%
Berat Panen	431	98.4%
Luas	0	0%
Suhu Air	0	0%
Oksigen Terlarut	0	0%
Salinitas	0	0%
pH	0	0%
Kolam	0	0%

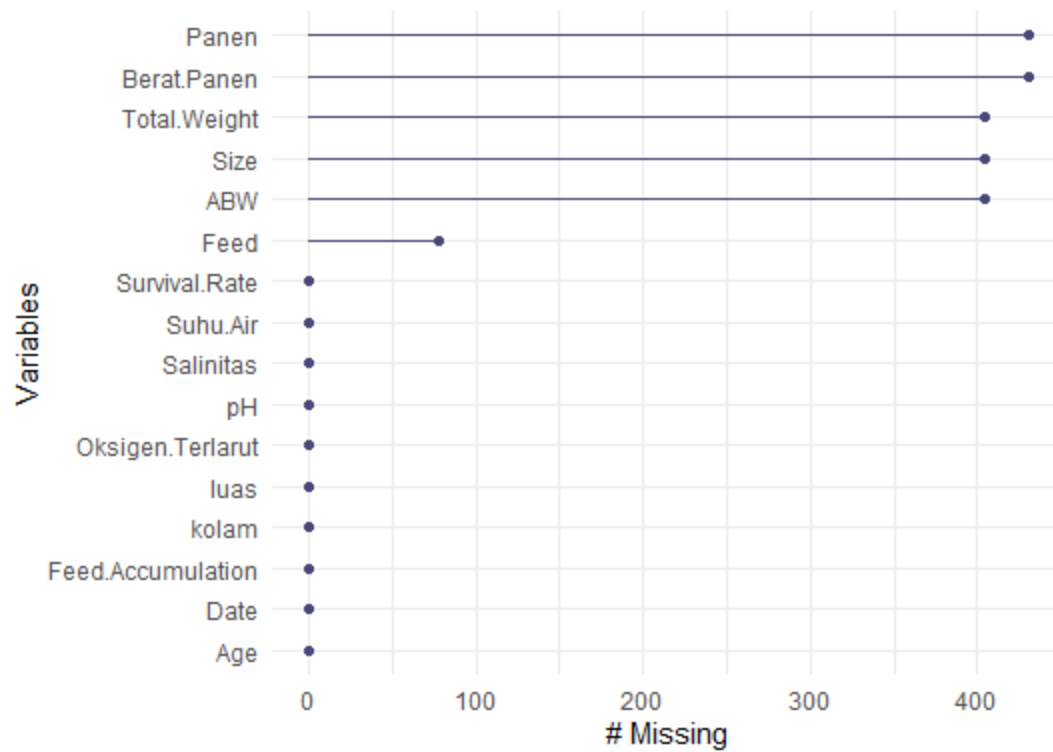


Figure 14 numbers of missing valur for each variable

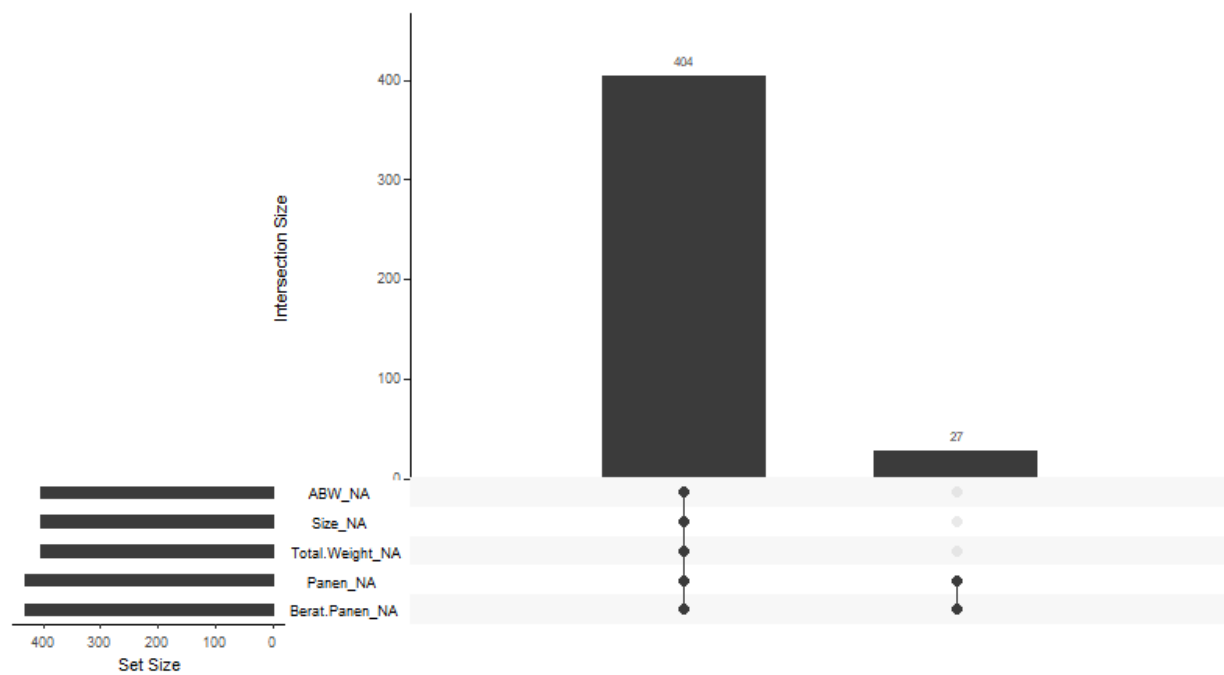


Figure 15 intersection missing value:

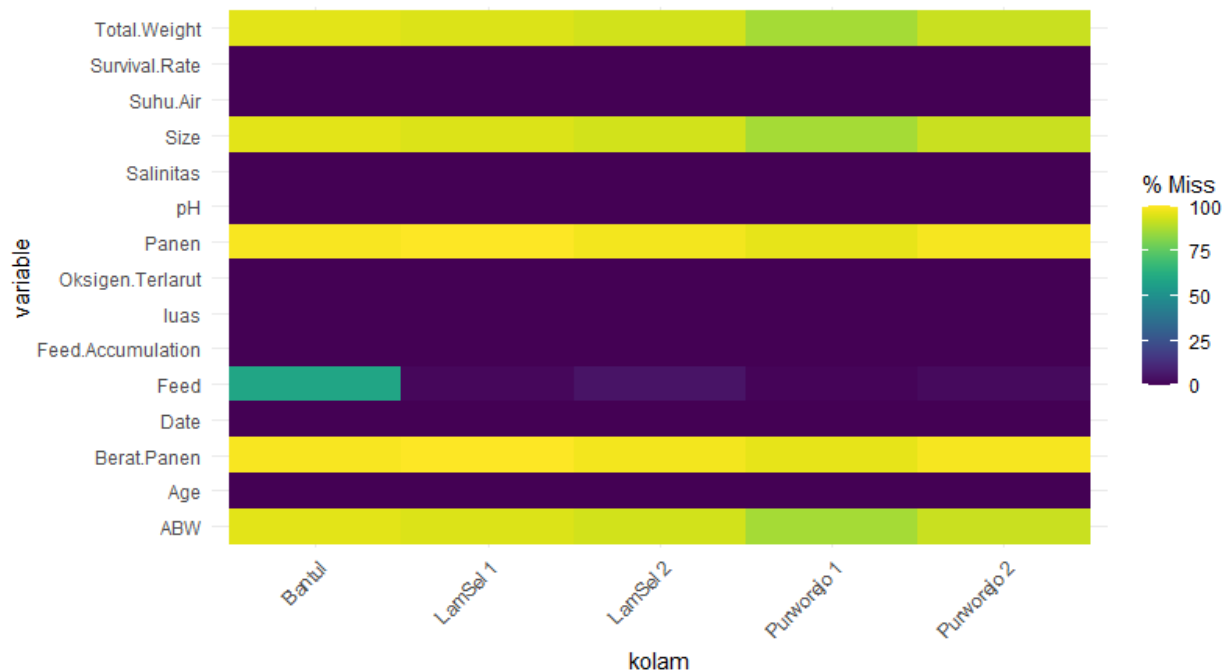


Figure 16 heatmap missing value

Dikarekankan begitu banyaknya missing value di beberapa variable seperti panen, berat panen, total weight, size, ABW, maka penulis (setelah berdiskusi dengan pihak Jala Tech) menganalisis menggunakan beberapa alternative pengukuran seperti productivity, survival rate, dan FCR.

Kolam	Productivity	Survival Rate	FCR
Bantul	8.18	81.2	1.05
Lampung Selatan 1	NA	105.0	NA
Lampung Selatan 2	115.0	106.0	1.17
Purworejo 1	27.7	82.9	5.90
Purworejo 2	27.8	55.5	2.22

b. Productivity

Productivity tiap kolam dapat diukur dengan mencari rasio antara biomassa yang dihasilkan saat panen dibandingkan dengan luas kolam. Sederhananya menggunakan rumus berikut

$$\text{Productivity} = \text{biomass} / \text{luas kolam}$$

Luas kolam digabungkan/dimerge menggunakan data siklus yang sebelumnya telah diolah. Pada Figure 17 menunjukkan bahwa Lampung Selatan 2 memiliki produktivitas tertinggi jika disbanding dengan 3 kolam lainnya yaitu Bantul, Purworejo 1 dan 2. Namun juga disayangkan bahwa data/ record pada kolam Lampung Selatan 1 tidak ada.

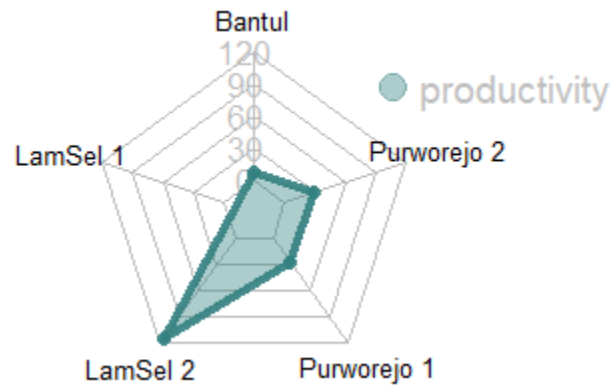


Figure 17 radarchart productivity for each place

c. Survival rate

Survival rate merupakan variable perkiraan presentasi udang yang masih bertahan hidup di kolam. Nampak bahwa kolam di lampung selatan 1 maupun Lampung Selatan 2 memiliki survival rate yang paling tinggi. Sehingga kedepannya dapat diperhitungkan untuk ekspansi ataupun pelebaran kolam.

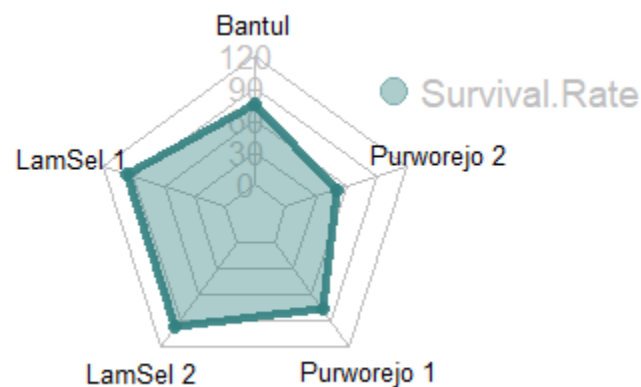


Figure 18 survival rate for each place

d. Feed Conversion Rate (FCR)

Feed Conversion Rate merupakan variable gabungan antara banyaknya pakan yang digunakan dibagi dengan berat total saat panen. Jika dilihat pada Figure 19, Kolam Purworejo memiliki FCR yang paling tinggi, sedangkan untuk Bantul dan Lampung Selatan 2 termasuk kolam dengan FCR paling kecil.

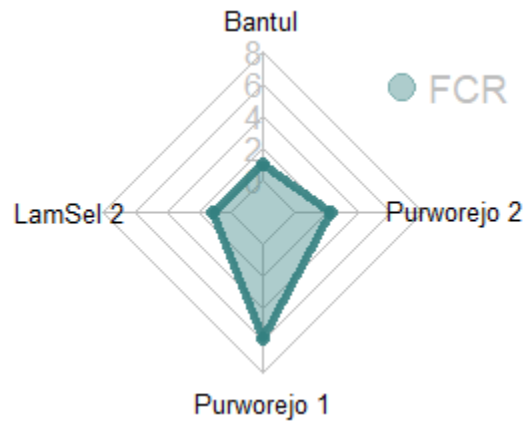


Figure 19 Feed Conversion Rate for each place

4. Kesimpulan

- a. Teridentifikasi belum adanya indikator jumlah benur yang harus disebar pada masing-masing kolam sebagai optimasi hasil panen
- b. kualitas suhu air yang paling baik terdapat pada kolam lampugn selatan, dan yang paling buruk di purworejo
- c. Kualitas oksigen terlarut yang paling jelek terdapat pada kolam purworejo 1
- d. Kualitas Salinitas perlu dievaluasi terdapat di kolam bantul dan kolam purworejo 1
- e. Kualitas pH tiap kolam telah memenuhi standard
- f. Data budidaya perlu ditingkatkan akurasi dan konsistensinya karena terdapat beberapa variable yang tingkat missing valuenya melebihi 90% dari record data yang ada
- g. Jika diperlukan dibuat dataset khusus untuk mengakomodir data pada saat panen, sehingga tidak mengganggu historical data saat proses pengembangan budidaya
- h. Data pada lampung selatan 1 tidak ditemukan hasil productivity dan FCR nya dikarenakan tidak tersedianya data
- i. Productivity kolam yang terbaik terdapat pada kolam lampung selatan 2
- j. Survival rate yang tertinggi terdapat pada kolam lampung selatan 1 dan 2
- k. FCR terendah terdapat pada kolam bantul dan lampung selatan 2, dan FCR tertinggi terdapat pada kolam Purworejo 1

