Friday, 25 October 2019

Vanilla Research- M.P.Appanna

Background:

Earlier this year, I decided to assume some responsibility over the managing of my family's coffee plantation. We are native to a place called Coorg, which is a forest hill station situated on the Western Ghats of South India, known for its coffee plantations. This was going to be the first year that the newly planted Vanilla vines, would flower.

Vanilla vines first flower typically 3 years after the vine cuttings are planted. The flowers have to be manually pollinated because the natural pollinator —a particular type of bee found in Mexico— is not native to India. Using a toothpick, I had to carefully carry out the procedure to bring about pollination. This took some getting used to; The middle column of the flower, an orchid, has to be delicately sliced so as to expose the reproductive organs. Then using the toothpick to carefully lift a flap containing pollen, the part above this flap is closed over it.

Successful pollination was still only a probability because it depended on the learning curve for the skill required to carry out the procedure. Labour is one of the main issues faced by most agriculturists in Coorg. Apart from the increasing cost of labour wages, there is also the challenge of varying quality of work. So for the especially delicate task of vanilla pollination, quality labour is difficult to come by. The optimal time for pollination is in the couple of hours around noon. Too early in the morning, and you may miss some new flowers. Too late, and some flowers might have started to close up. Since there were only a manageable number of vines which were blossoming in my plantation, I decided to take on the task of pollinating this, myself.

The flowering vanilla vine locations were distributed randomly in a grid of various trees. The vines are planted such that they climb the tree beside it. The safest way to observe for any new flowering vines while also pollinating all flowers for that day, was to make sure that I walked through the grid in an orderly manner. This order I hoped, would enable me to become more intuitive with the regular sequence of the vines, allowing me the possibility to observe some otherwise redundant details. Also, an orderly manner serves as a way of tracking progress for the day's work. Motivation is always useful.

Going to each of the vines, pollinating the flowers and then noting down the data parameters for the day, was quite a tedious task. But this sample can be useful in identifying certain tendencies applicable to Vanilla flowering. You may refer the Jupyter notebook for the data cleaning and visualisation.

Friday, 25 October 2019

Abstract:

I decided to create a journal on the Notes app of my phone, detailing insights about each of the newly flowering vines. For this, I needed to assign a key (unique reference) to indicate each of the locations. Seeing as how the locations might be affected by the onset of new flowering locations (new vines), and casualties due to trees falling, a plain sequential numbering system would not suffice. I observed that the tree grid consisted of mainly 3 types of trees. The Adke (arekenut*),* Silver Oak (grevillea) and small tree stumps. The stumps are planted in empty spots in the tree grid and is of a tree variety which grows quite easily (like *daddops*). The columns are easily identifiable and make up the first identifier of the key code. The second identifier is a short hand way of expressing the particular tree that the vine is attached to. The trees serve as quick visual reference points to indicate the location for a vine. For example:

Key	**1st Identifier (Column)**	**2nd Identifier (Tree code)**	**2nd identifier description**
11*a	11	a	On the 1st Adke in the 11th column
7*3a3s	7	3a3s	On the 3rd stump 's' which comes after the 3rd Adke 'a'
4*2a	4	2a	On the 2nd Adke 'a'
4*3a2g	4	3a2g	On the gravilia 'g' which comes after an Adke 'a'

Next, I had to select the data parameters, making sure to not have so many that it would be impractical to manually quantify. At the same time there needed to be enough to allow me to gain some insights. The below chart details some changes made in the selection of parameters to populate, as well as the 'purpose definition' of the 2nd Iteration.

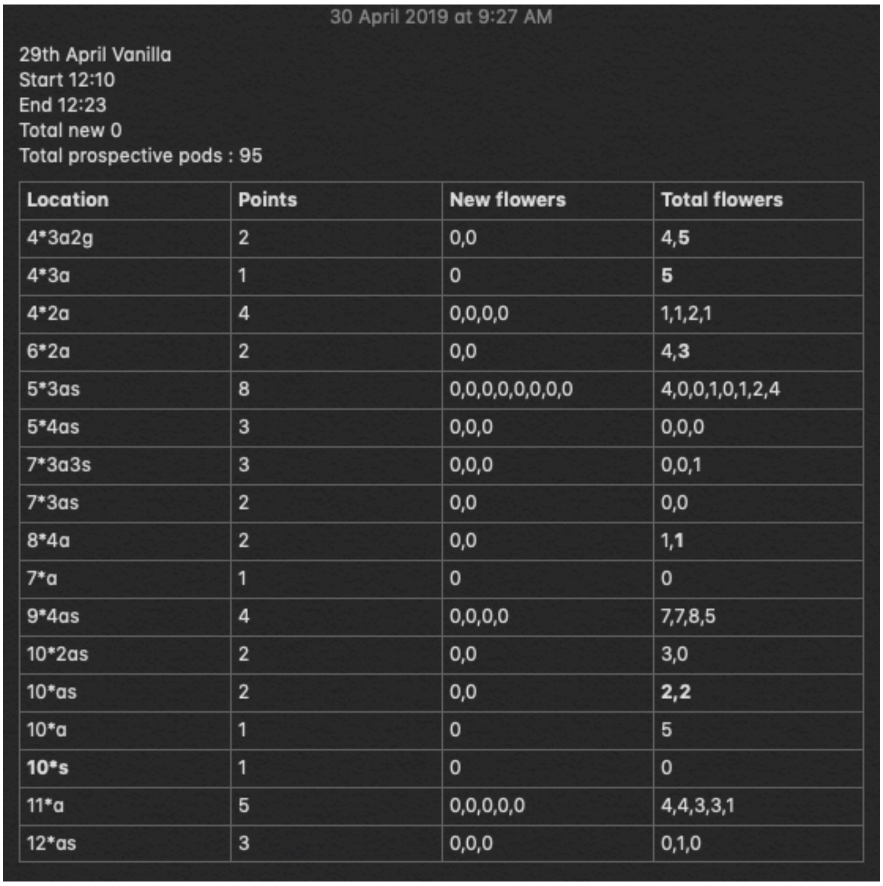
1st Iteration	**2nd Iteration**	**Purpose**
Column and Row	Location	Identify the vine
Points	Points	Track number of flowering points at a location
Flower today 10-12 am	New Flowers	Number of new flowers to be pollinated
Flower Total	Flower Total	Potential crop
Empty	-	-
Budding	-	-

*Definitions:

Flower Total - If a flower is successfully pollinated, the petals of the flower typically stay attached to the stem section, which then becomes the vanilla pod. If the entire flower withers and falls off either the next day or in 2 days, then it is an indicator that the pollination was unsuccessful.

Data Collecting and Cleaning:

1. I used the Notes app in the IPhone, to manually enter day wise data. The values of the parameters were entered inside a tabular.



30 April 2019 at 9:27 AM

29th April Vanilla
Start 12:10
End 12:23
Total new 0
Total prospective pods : 95

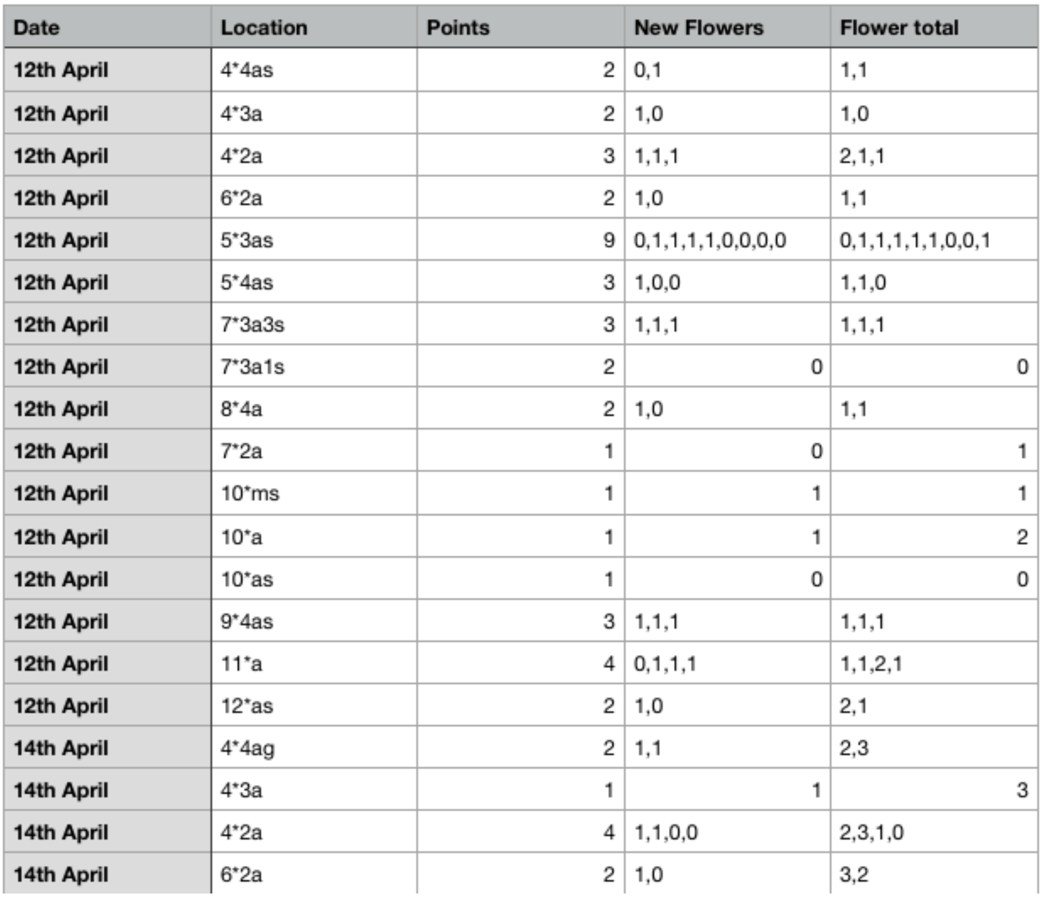
Location	Points	New flowers	Total flowers
4*3a2g	2	0,0	4,**5**
4*3a	1	0	**5**
4*2a	4	0,0,0,0	1,1,2,1
6*2a	2	0,0	4,**3**
5*3as	8	0,0,0,0,0,0,0,0	4,0,0,1,0,1,2,4
5*4as	3	0,0,0	0,0,0
7*3a3s	3	0,0,0	0,0,1
7*3as	2	0,0	0,0
8*4a	2	0,0	1,**1**
7*a	1	0	0
9*4as	4	0,0,0,0	7,7,8,5
10*2as	2	0,0	3,0
10*as	2	0,0	**2,2**
10*a	1	0	5
10*s	1	0	0
11*a	5	0,0,0,0,0	4,4,3,3,1
12*as	3	0,0,0	0,1,0

- The order of the locations is indicative of the actual walking path while observing the vines. A walking path goes between 2 columns. Any flowers encountered on the path is identified as a flowering vine and is tracked.

- The last character in the ‘Location’ code is indicative of what kind of tree the vine is climbing. ’s’ for stump, ‘g’ for gravilia and ‘a’ for arekenut.

2. Next I manually copied the tabular in Notes for the different days and made a continuous table of date-wise values in the Mac spreadsheet software “Numbers”. This was then exported into a .csv file which I could import into Jupyter Notebooks.

Table 1



Date	Location	Points	New Flowers	Flower total
12th April	4*4as	2	0,1	1,1
12th April	4*3a	2	1,0	1,0
12th April	4*2a	3	1,1,1	2,1,1
12th April	6*2a	2	1,0	1,1
12th April	5*3as	9	0,1,1,1,1,0,0,0,0	0,1,1,1,1,1,0,0,1
12th April	5*4as	3	1,0,0	1,1,0
12th April	7*3a3s	3	1,1,1	1,1,1
12th April	7*3a1s	2	0	0
12th April	8*4a	2	1,0	1,1
12th April	7*2a	1	0	1
12th April	10*ms	1	1	1
12th April	10*a	1	1	2
12th April	10*as	1	0	0
12th April	9*4as	3	1,1,1	1,1,1
12th April	11*a	4	0,1,1,1	1,1,2,1
12th April	12*as	2	1,0	2,1
14th April	4*4ag	2	1,1	2,3
14th April	4*3a	1	1	3
14th April	4*2a	4	1,1,0,0	2,3,1,0
14th April	6*2a	2	1,0	3,2

3. Within Jupyter Notebooks I attempted to carry out all further data modification operations.

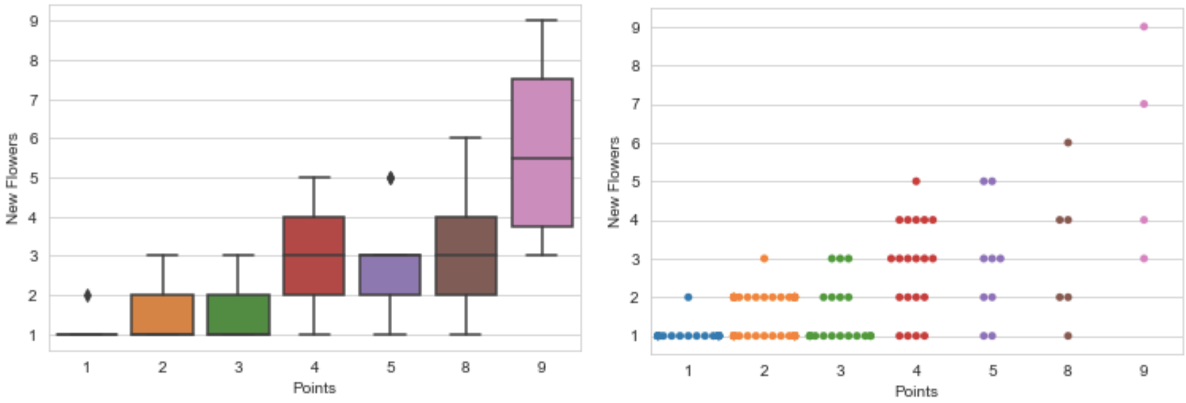
   - The 'New Flowers’ and ‘Flower total’ columns had comma separate values for each flower point. This had to be exploded as a line item each.
   - As a consequence of attempting to carry out the above operation, I had to use a method to identify a mismatch in length of csv values of the 2 columns, so that I did not get an ‘arrays must all be same length’ error, while chaining the individual elements in separate rows.

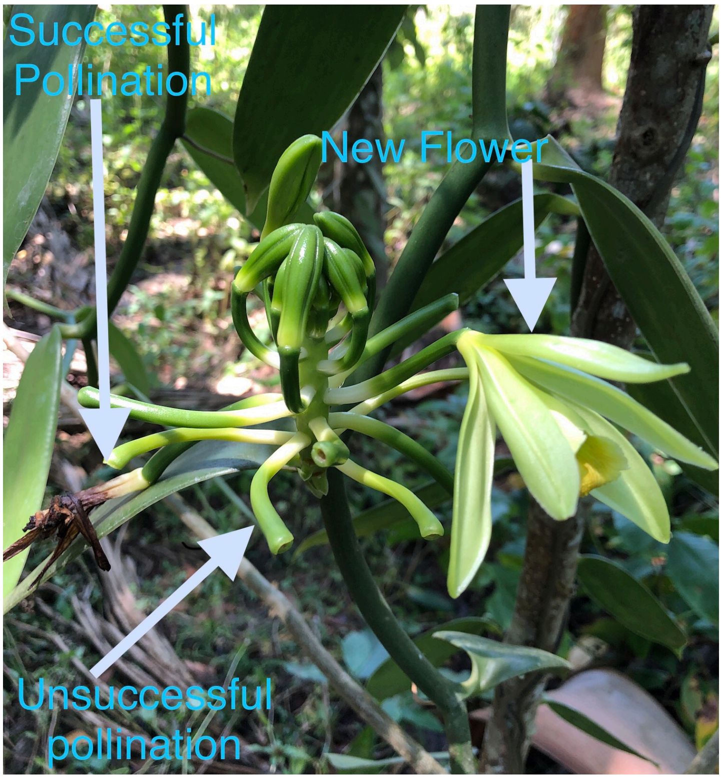
- The mismatched rows in the data frame were due to typos during data entry which had to be debugged by manual over write of the cells.
- The cleaned table was then exported to a csv file which could then be used in the data visualisation notebook.

Visualisation

How many new flowers can be expected for the season?

How to plan out optimal pollination day strategy?



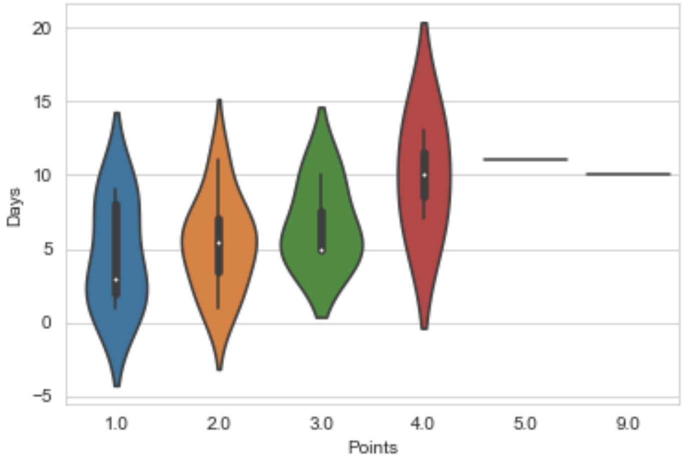


The above charts represent the number of 'New Flowers' on a day, at a vine location. The vine location is represented by the number of flowering points it has. The part on the vine from where the flowers are blossoming are termed as 'Points'. A certain point keeps flowering for a couple of days, till it exhausts its potential. To the left is a picture of 1 flowering point with its different stages.

The box plot can be used to infer the inter-quartile ranges.

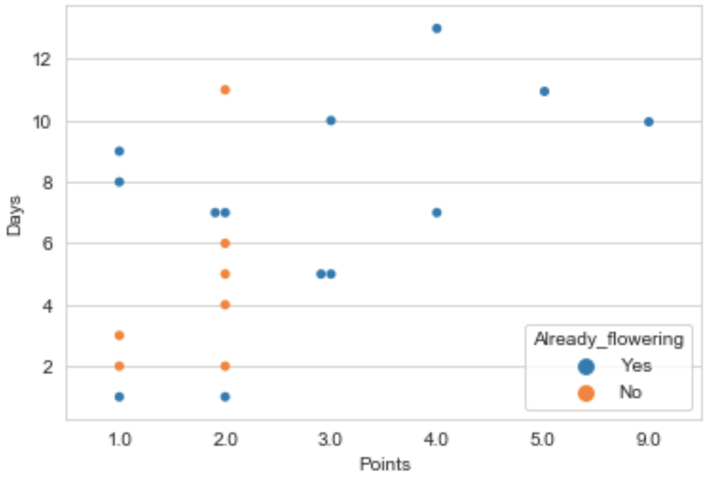
Each dot in the swarm plot represents occurrence of flowers at a location on a particular day.The location has been further categorised by 'Points' since different locations may have different number of flowering points. The assumption is that the number of flowering points may have significant affect on selecting an aggregate number of New flowers for a location.

Number of New flowers for the season can be derived by multiplying the median number of New flowers (based on number of Points) with total number of days of flowering. The below chart can be used to determine how many days to factor for a particular 'Point'.



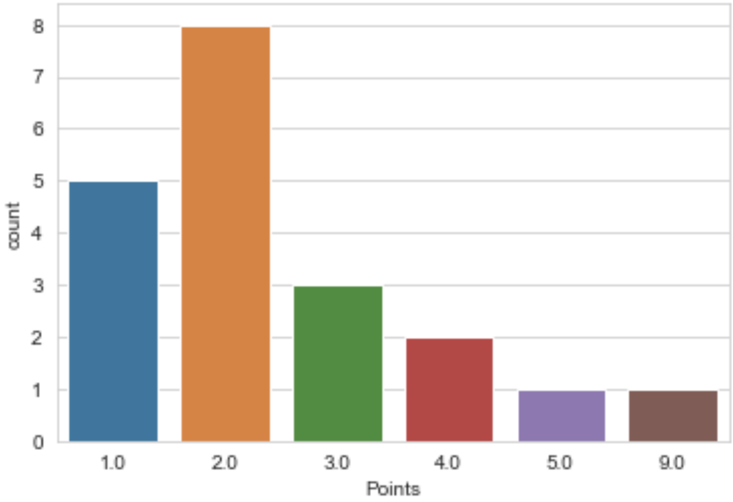
From the violin plot above, we can infer inter-quartile information about the Days. The vertical black rectangle represents 50% of the data between the 25th and 75th percentile. The white dot represents the median.

The median can be a fairly reliable number to use, but we must be aware of other factors which might affect it. For instance, when the collection of this data set and manual pollination began, there were some vines which had not yet started flowering. If these 'late bloomers' affect your selection of the median for days, the below swarm plot may offer an explanation. My observation is that, since the days are uniformly and widely distributed, in both cases, I would favour quantity of data, and thus select the median.



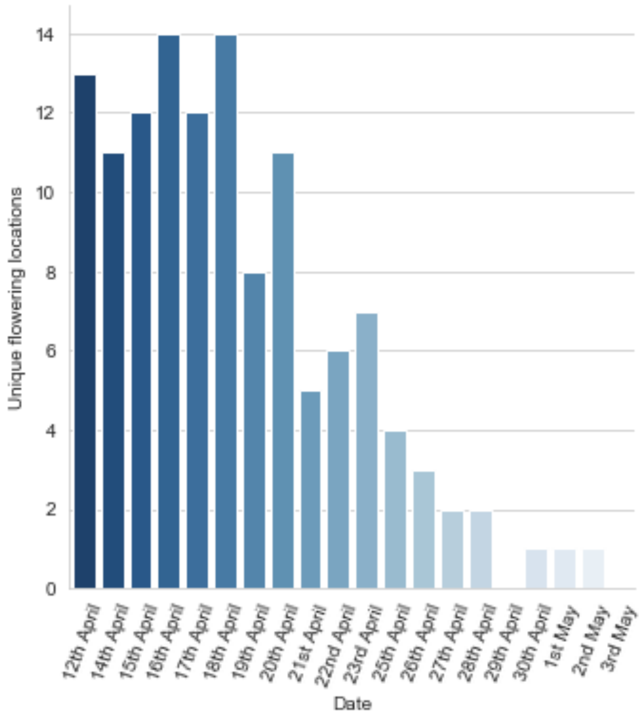
Each dot in the above swarm represents each of the locations. We can see that most locations are of the 2 Point type. It appears that no location has flowered for more than 13 days. However since this vine was already flowering, there may have been more than 13 days of flowering.

As the number of points increases, the number of such locations is rarer but it typically flowers for more days.



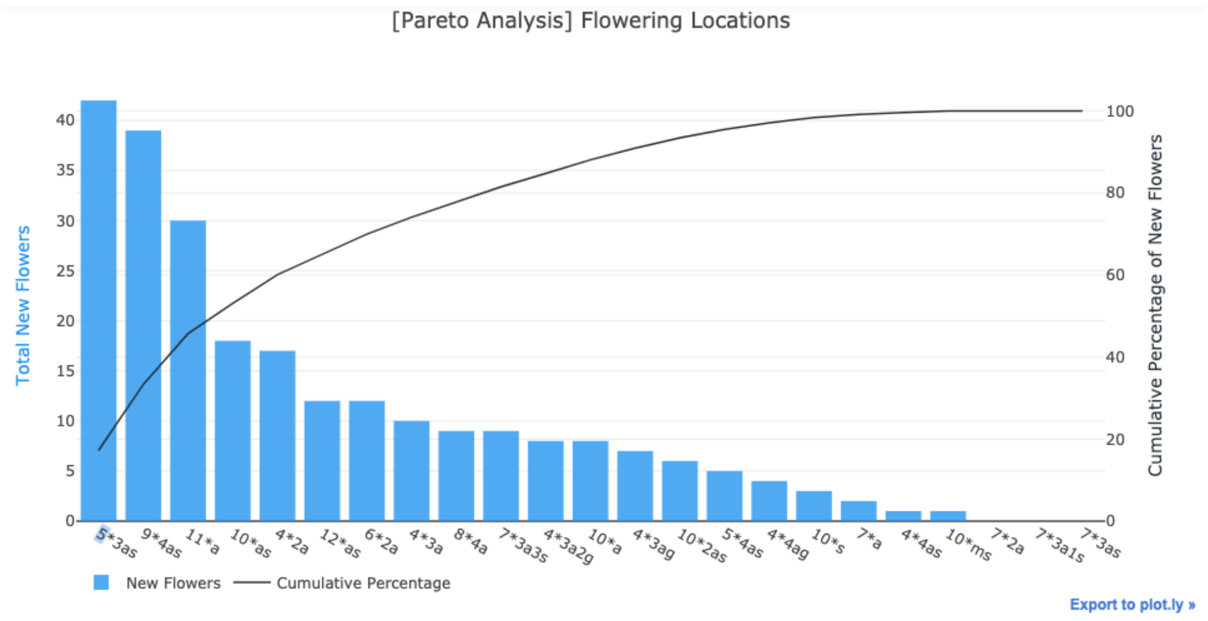
The distribution of count of locations as per the number of points, can be visualised by the above count plot. As we can see, the 2-point vine locations form the majority with 8 out of the total of 20 locations.

The probability of occurrence of flowers at a location, on a particular day, can add another element of decision making while selecting which vines to attempt pollination for each day. Since the data collection started when most locations were already flowering, the bar chart below shows how the number of unique locations gradually reduce towards the end of the season. It may be fair to assume a normal distribution. Highest number of locations on a given day was 14 and it took 12 days to reach the least number of flowering locations which was 1.

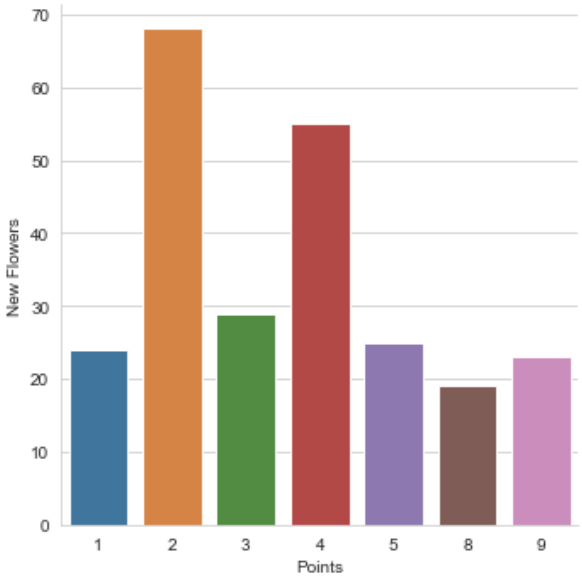


To plan pollination strategy we may take the total of estimated New flowers for the season, ordered by highest to lowest, and then attempt to pollinate the top 80% which is likely to be caused by 20% of the locations. The hope is that the results would tend towards this rather ideal distribution with more data.

Practically however, given the limited locations of this dataset, it appears that 80% of the New flowers were from 50% (10 out of 20) of the locations. This still results in close to 50% reduction in labour time if we assume that the total time spent varies most significantly by number of locations. As an aside, in the population of this data, each location meant that I had to manually key in the numbers on my note, so there was a significant fixed time cost (the actual pollination was faster) associated with more locations.



Total number of new flowers per type of location . What multipoint location gave most number of new flowers, for the entire season?



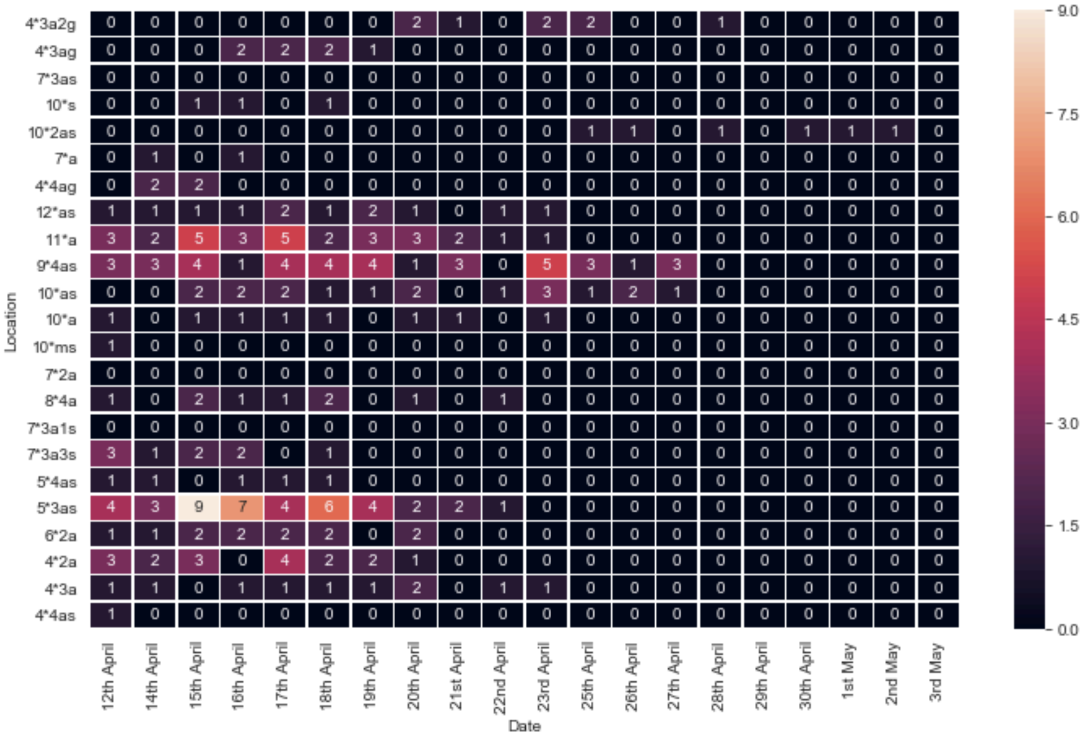
The 2 point and 4 point vines seem to be most productive in new flowers whereas all the other categories are closer to each other despite having different median days of flowering and number of flowers.

When we dig deeper, we find that there are only 2 locations which are responsible for all the flowers in the 4 point category. And there are 8 locations responsible for the flowers in the 2 point category.

The 4 point locations are : 4*2a and 9*4as

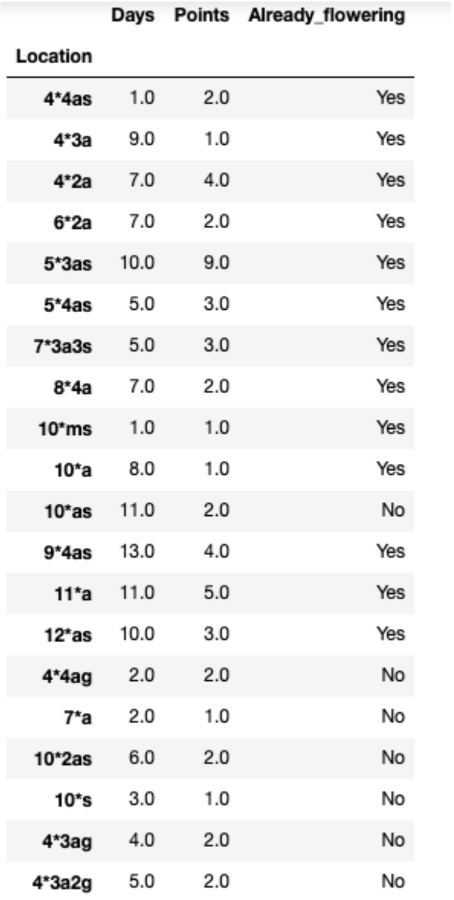
The 2 point locations are : 4*4as, 6*2a, 8*4a, 10*as, 4*4ag, 10*2as, 4*3ag, 4*3a2g

We can use the heat map below to observe the performance of these outlier locations across the season in terms of new flowers per day

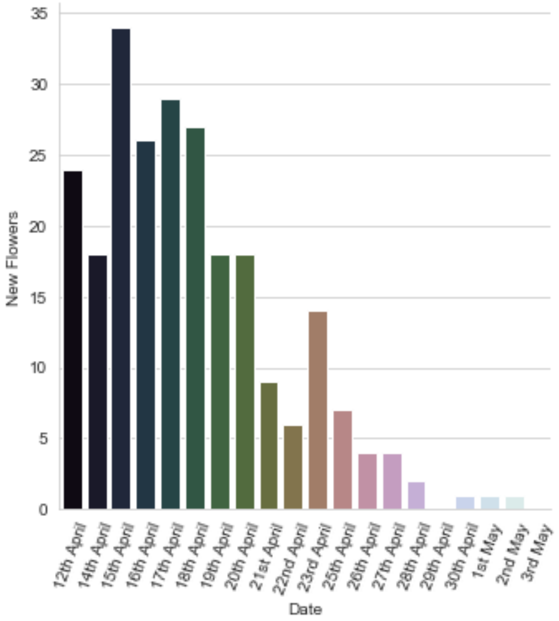


Location	12th April	14th April	15th April	16th April	17th April	18th April	19th April	20th April	21st April	22nd April	23rd April	25th April	26th April	27th April	28th April	29th April	30th April	1st May	2nd May	3rd May
4*3a2g	0	0	0	0	0	0	0	2	1	0	2	2	0	0	1	0	0	0	0	0
4*3ag	0	0	0	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7*3as	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10*s	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10*2as	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	1	0
7*a	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4*4ag	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12*as	1	1	1	1	2	1	2	1	0	1	1	0	0	0	0	0	0	0	0	0
11*a	3	2	5	3	5	2	3	3	2	1	1	0	0	0	0	0	0	0	0	0
9*4as	3	3	4	1	4	4	4	1	3	0	5	3	1	3	0	0	0	0	0	0
10*as	0	0	2	2	2	1	1	2	0	1	3	1	2	1	0	0	0	0	0	0
10*a	1	0	1	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0
10*ms	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7*2a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8*4a	1	0	2	1	1	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0
7*3a1s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7*3a3s	3	1	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5*4as	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5*3as	4	3	9	7	4	6	4	2	2	1	0	0	0	0	0	0	0	0	0	0
6*2a	1	1	2	2	2	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0
4*2a	3	2	3	0	4	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
4*3a	1	1	0	1	1	1	1	2	0	1	1	0	0	0	0	0	0	0	0	0
4*4as	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The maximum number of new flowers are from the location 5*3as which has the highest number of flowering points, 9. Also since this vine was already flowering, there may have been more flowers.



	Days	Points	Already_flowering
Location			
4*4as	1.0	2.0	Yes
4*3a	9.0	1.0	Yes
4*2a	7.0	4.0	Yes
6*2a	7.0	2.0	Yes
5*3as	10.0	9.0	Yes
5*4as	5.0	3.0	Yes
7*3a3s	5.0	3.0	Yes
8*4a	7.0	2.0	Yes
10*ms	1.0	1.0	Yes
10*a	8.0	1.0	Yes
10*as	11.0	2.0	No
9*4as	13.0	4.0	Yes
11*a	11.0	5.0	Yes
12*as	10.0	3.0	Yes
4*4ag	2.0	2.0	No
7*a	2.0	1.0	No
10*2as	6.0	2.0	No
10*s	3.0	1.0	No
4*3ag	4.0	2.0	No
4*3a2g	5.0	2.0	No



The chart above shows maximum number of new flowers on any given day.

The table on the left gives details about the number of points of a particular location and the number of days of flowering.

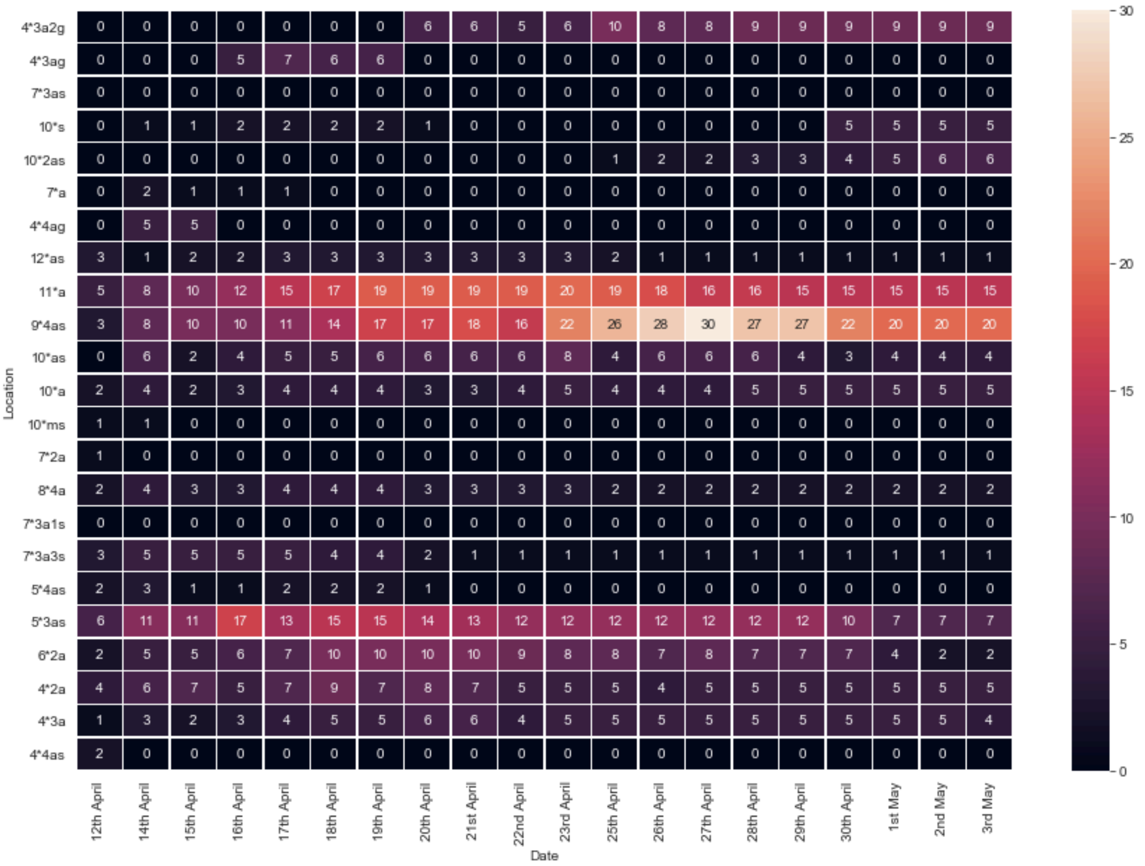
As seen in the heat map above, the maximum number of new flowers are from the location with highest number of flowering points, 9. Also since this vine was already flowering, there may have been more flowers.

Friday, 25 October 2019

What was the trend of successful pollinations across the season? - Flower Total

The below heat map gives an idea of the potentially successful pollinations at a location on a given day. This is a total number of flowers and includes new flowers and old flowers. Out of the old flowers if the flower stock remains for more than about 2 days, it generally stays on, indicating a successful pollination.

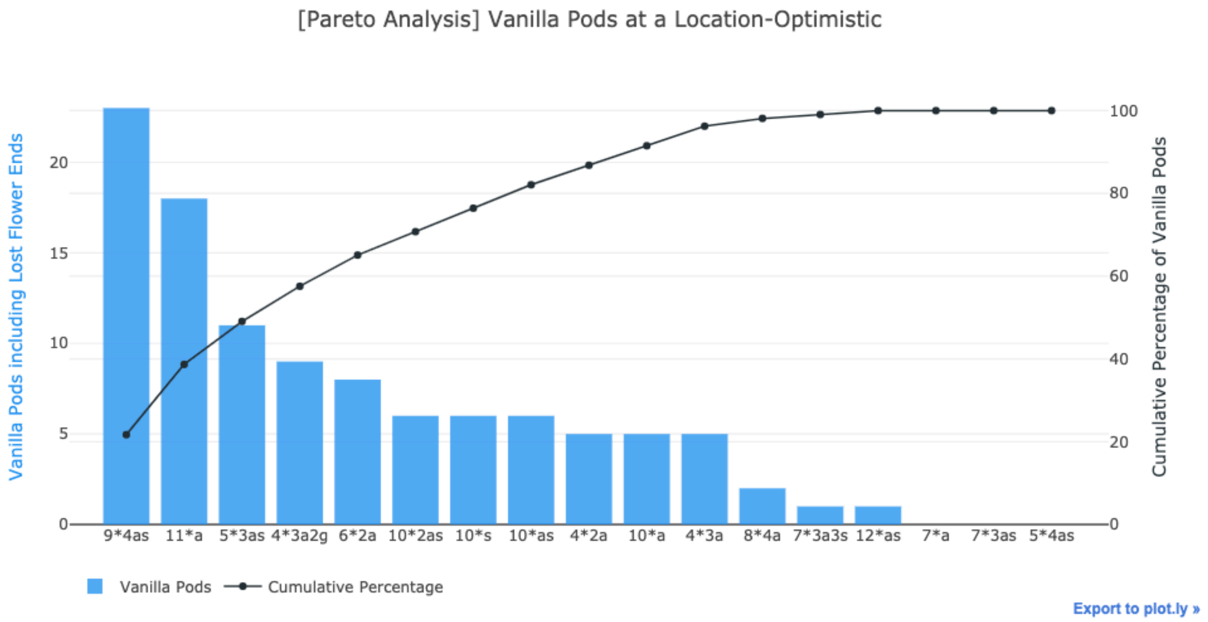
Further drops in day to day numbers may be caused by pests like caterpillars and ants which tend to eat the leftover flowers. The number of total flowers on the last day ,3rd May, is indicative of the expected yield for the season.



Location	12th April	14th April	15th April	16th April	17th April	18th April	19th April	20th April	21st April	22nd April	23rd April	25th April	26th April	27th April	28th April	29th April	30th April	1st May	2nd May	3rd May
4*3a2g	0	0	0	0	0	0	0	6	6	5	6	10	8	8	9	9	9	9	9	9
4*3ag	0	0	0	5	7	6	6	0	0	0	0	0	0	0	0	0	0	0	0	0
7*3as	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10*s	0	1	1	2	2	2	2	1	0	0	0	0	0	0	0	0	5	5	5	5
10*2as	0	0	0	0	0	0	0	0	0	0	0	1	2	2	3	3	4	5	6	6
7*a	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4*4ag	0	5	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12*as	3	1	2	2	3	3	3	3	3	3	3	2	1	1	1	1	1	1	1	1
11*a	5	8	10	12	15	17	19	19	19	19	20	19	18	16	16	15	15	15	15	15
9*4as	3	8	10	10	11	14	17	17	18	16	22	26	28	30	27	27	22	20	20	20
10*as	0	6	2	4	5	5	6	6	6	6	8	4	6	6	6	4	3	4	4	4
10*a	2	4	2	3	4	4	4	3	3	4	5	4	4	4	5	5	5	5	5	5
10*ms	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7*2a	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8*4a	2	4	3	3	4	4	4	3	3	3	3	2	2	2	2	2	2	2	2	2
7*3a1s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7*3a3s	3	5	5	5	5	4	4	2	1	1	1	1	1	1	1	1	1	1	1	1
5*4as	2	3	1	1	2	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0
5*3as	6	11	11	17	13	15	15	14	13	12	12	12	12	12	12	12	10	7	7	7
6*2a	2	5	5	6	7	10	10	10	10	9	8	8	7	8	7	7	7	4	2	2
4*2a	4	6	7	5	7	9	7	8	7	5	5	5	4	5	5	5	5	5	5	5
4*3a	1	3	2	3	4	5	5	6	6	4	5	5	5	5	5	5	5	5	5	4
4*4as	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

There were some vanilla pods developing without the trailing flower part which was eaten by an insect like a caterpillar; this is not included in the final count for 3rd May. The difference can be factored as damage due to insects.

The Pareto chart below gives the total yield per location including the damaged vanilla pods. You can see how some locations are responsible for most of the New Flowers. 8 locations are holding 80% of the produce.



The vine with the most number of vanilla pods (most costly one) appears to be location 9*4as which is a 4 point vine holding about 23 pods.

Overall the success rate of the pollination is the fraction of total number of vanilla pods to the total number of new flowers, across the season.

New Flowers total is 243, Successful pollinations 106, Expected yield 86.

Success Rate of pollination: 43.6%