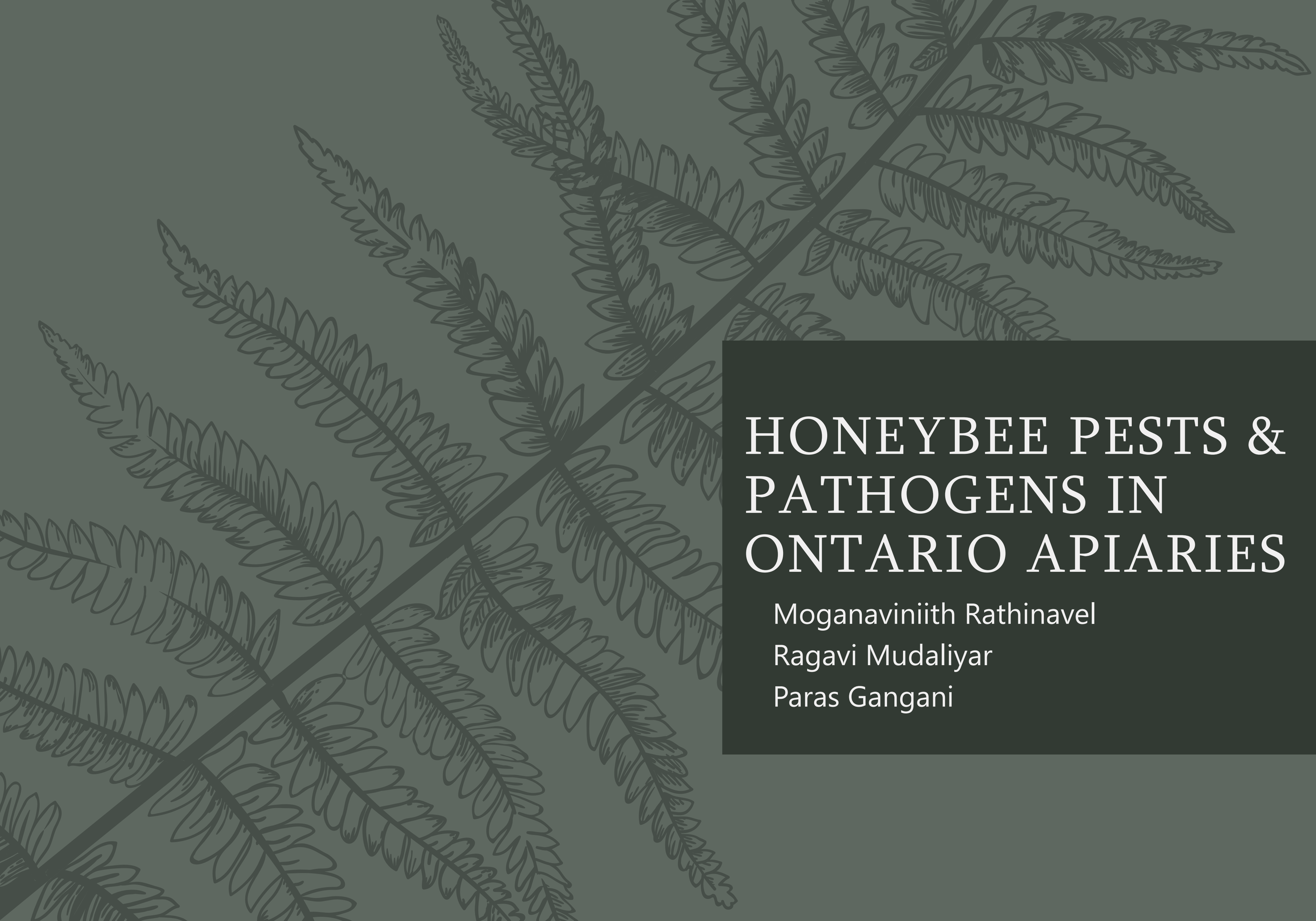
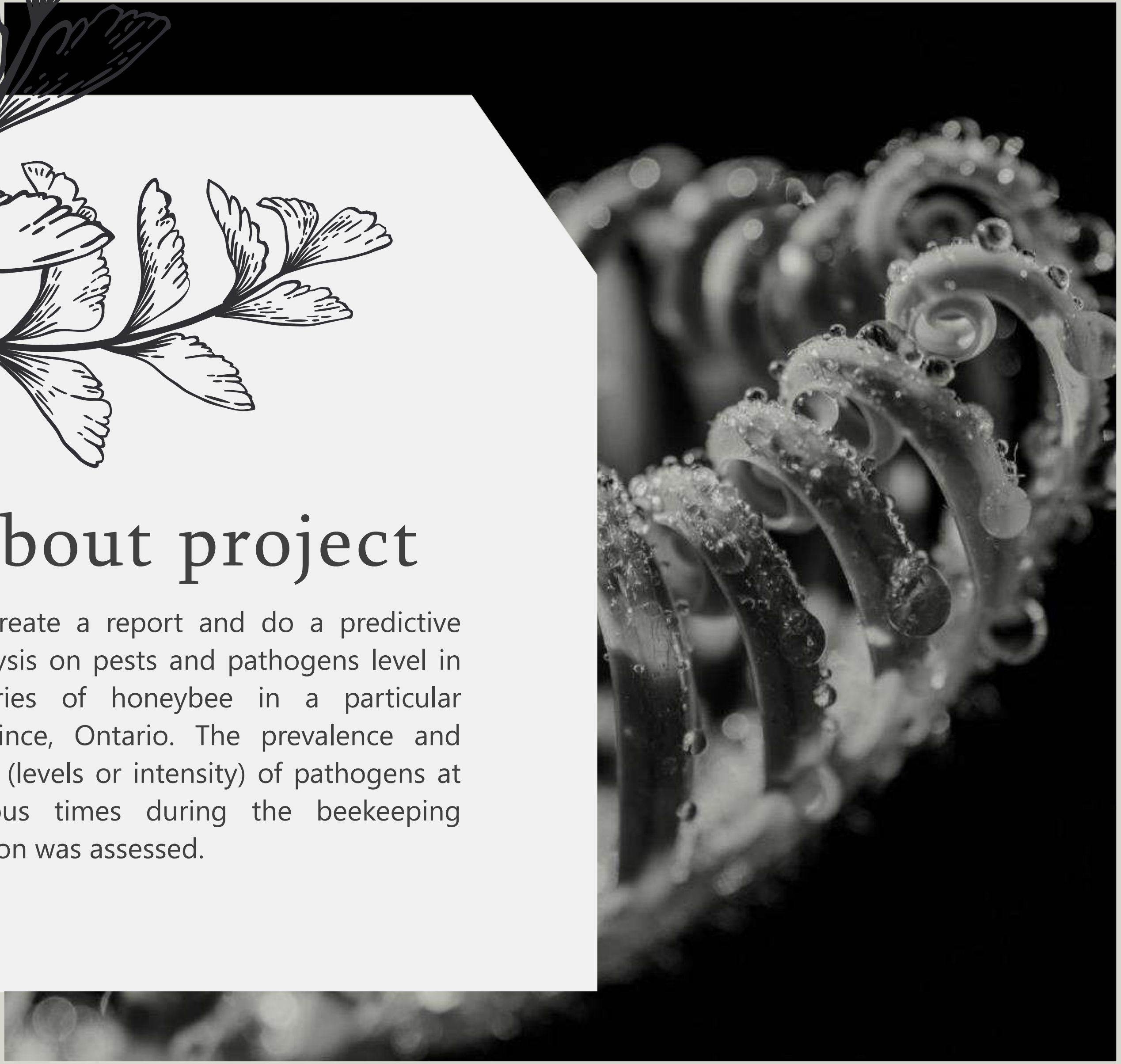


HONEYBEE PESTS & PATHOGENS IN ONTARIO APIARIES

Moganaviniith Rathinavel
Ragavi Mudaliyar
Paras Gangani

Table of Contents

1) About project
2) Meet our Team
3) Project Status
4) Project Phase
5) Milestones
6) About the Dataset
7) Documentation
8) Datatypes of variables and missing values distribution for year 2019
9) Cleaning Dataset - 2019
10) Dataset correlation - 2019
11) Dataset visualization - 2019
12) Dataset visualization - 2019
13) Predictive analysis - K means clustering
14) Predictive analysis - K means clustering
15) Project Cost
16) Issues or Challenges encountered this week and what was done to overcome them
17) Communications
18) Team meetings
19) Activities Completed This week
20) Activities to be Completed Before Next Report
21) Thank you



About project

To create a report and do a predictive analysis on pests and pathogens level in apiaries of honeybee in a particular province, Ontario. The prevalence and load (levels or intensity) of pathogens at various times during the beekeeping season was assessed.

Meet our Team

Rajalakshmi Nagarajan

Team lead/Developer

Ragavi Mudaliyar

Communicator

Paras Gangani

Analyst

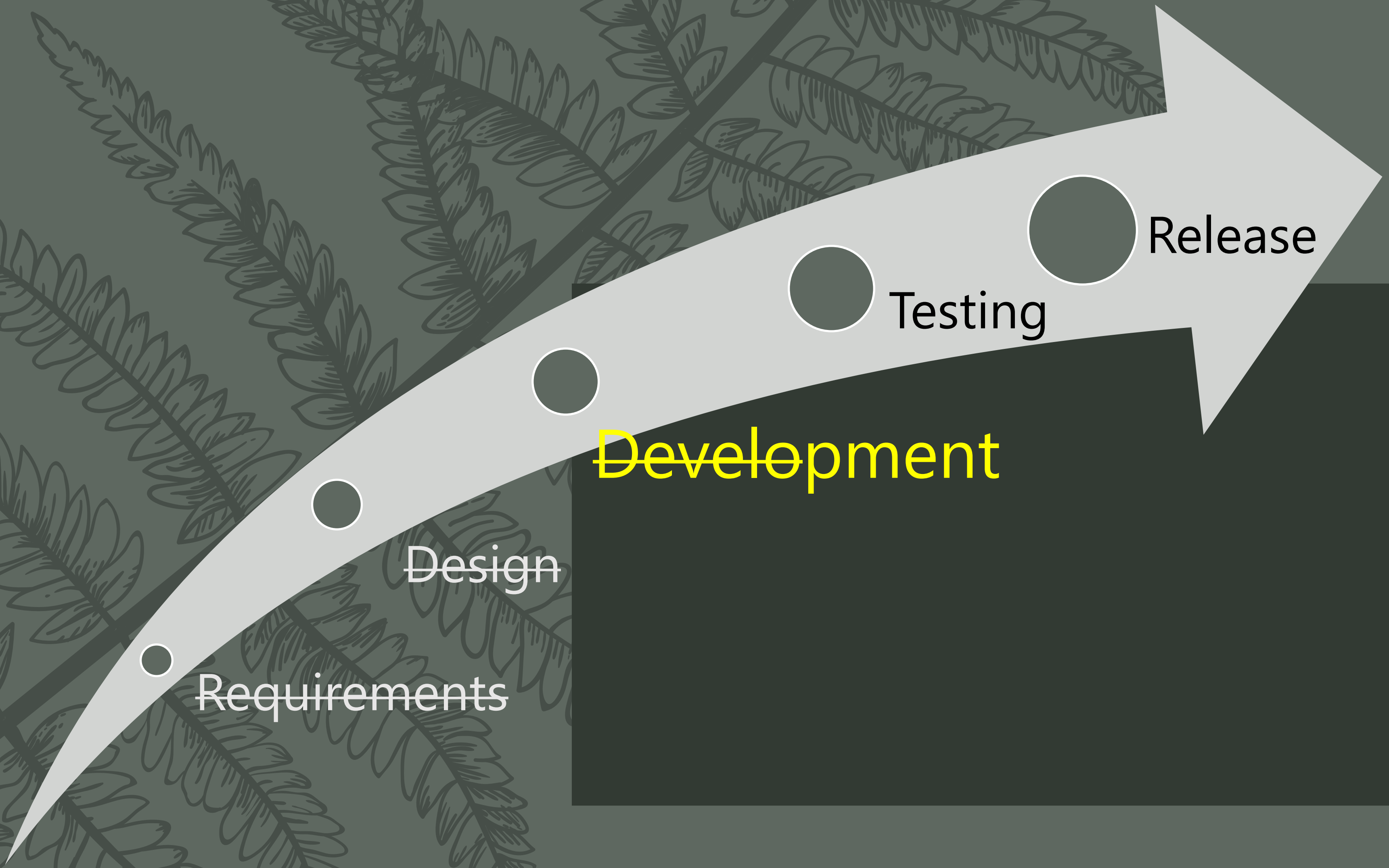
Project Status

✓ On Track

Date of report: 20th February. 2023

Date of last report: 15th February. 2023

Project Phase



~~Requirements~~

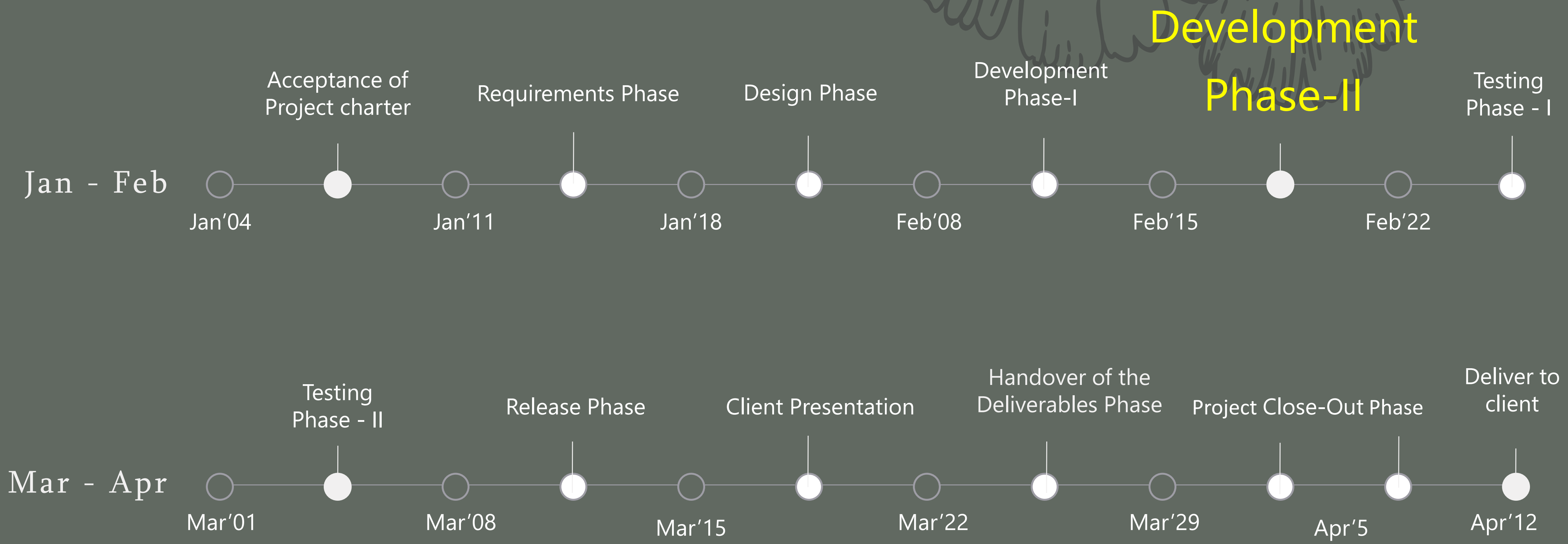
~~Design~~

~~Develo~~pment

Testing

Release

Milestones



Jan - Feb

Jan’04 – Acceptance of Project charter – Jan’11 – Requirements Phase – Jan’18 – Design Phase – Feb’08 – Development Phase-I – Feb’15 – Development Phase-II – Feb’22 – Testing Phase - I

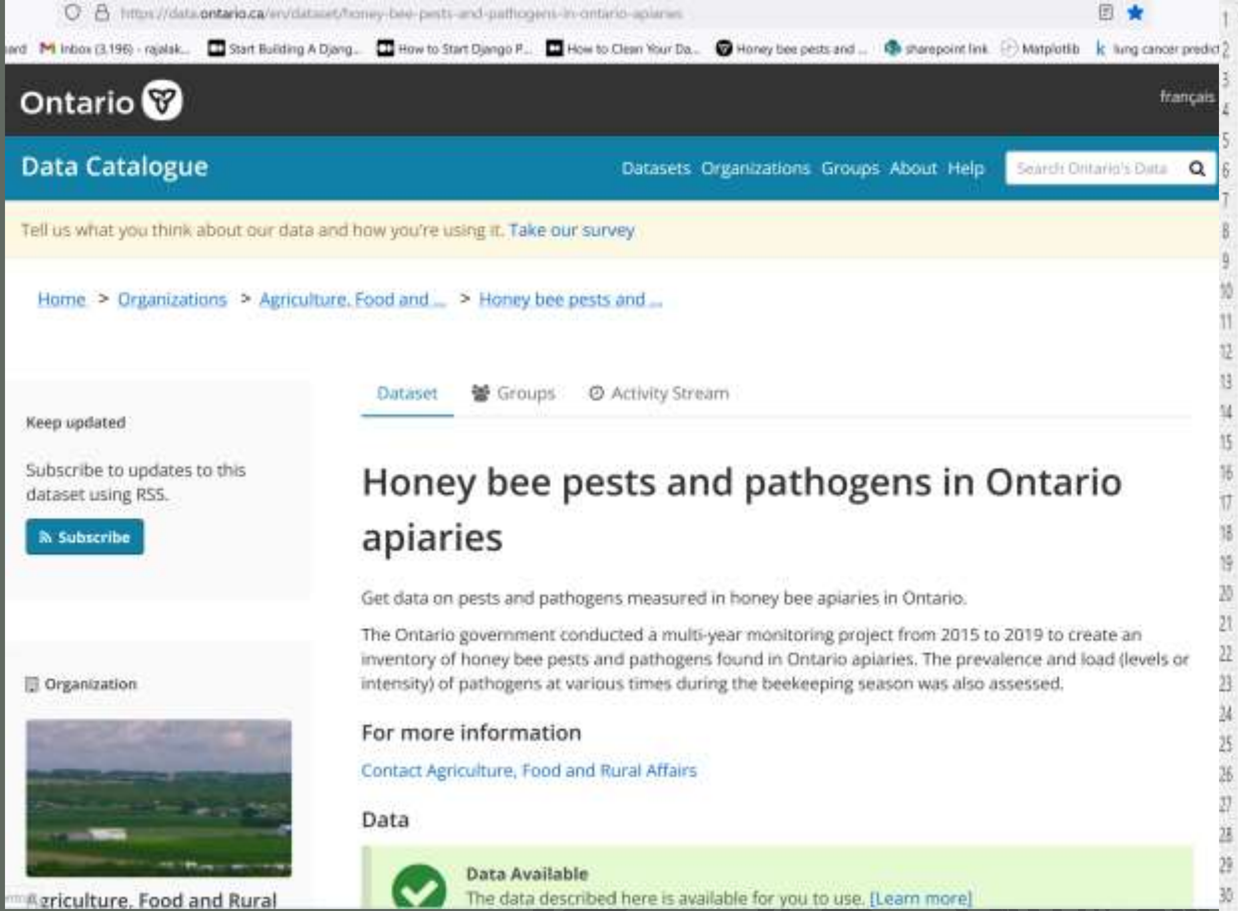
Mar - Apr

Mar’01 – Testing Phase - II – Mar’08 – Release Phase – Mar’15 – Client Presentation – Mar’22 – Handover of the Deliverables Phase – Mar’29 – Project Close-Out Phase – Apr’5 – Deliver to client – Apr’12

1 Requirements

About the Dataset

Website reference: https://data.ontario.ca/en/dataset/honey-bee-pests-and-pathogens-in-ontario-apiaries



https://data.ontario.ca/en/dataset/honey-bee-pests-and-pathogens-in-ontario-apiaries

Tell us what you think about our data and how you're using it. Take our survey

Home > Organizations > Agriculture, Food and ... > Honey bee pests and ...

Keep updated

Subscribe to updates to this dataset using RSS.

Organization

Agriculture, Food and Rural

Honey bee pests and pathogens in Ontario apiaries

Get data on pests and pathogens measured in honey bee apiaries in Ontario.

The Ontario government conducted a multi-year monitoring project from 2015 to 2019 to create an inventory of honey bee pests and pathogens found in Ontario apiaries. The prevalence and load (levels or intensity) of pathogens at various times during the beekeeping season was also assessed.

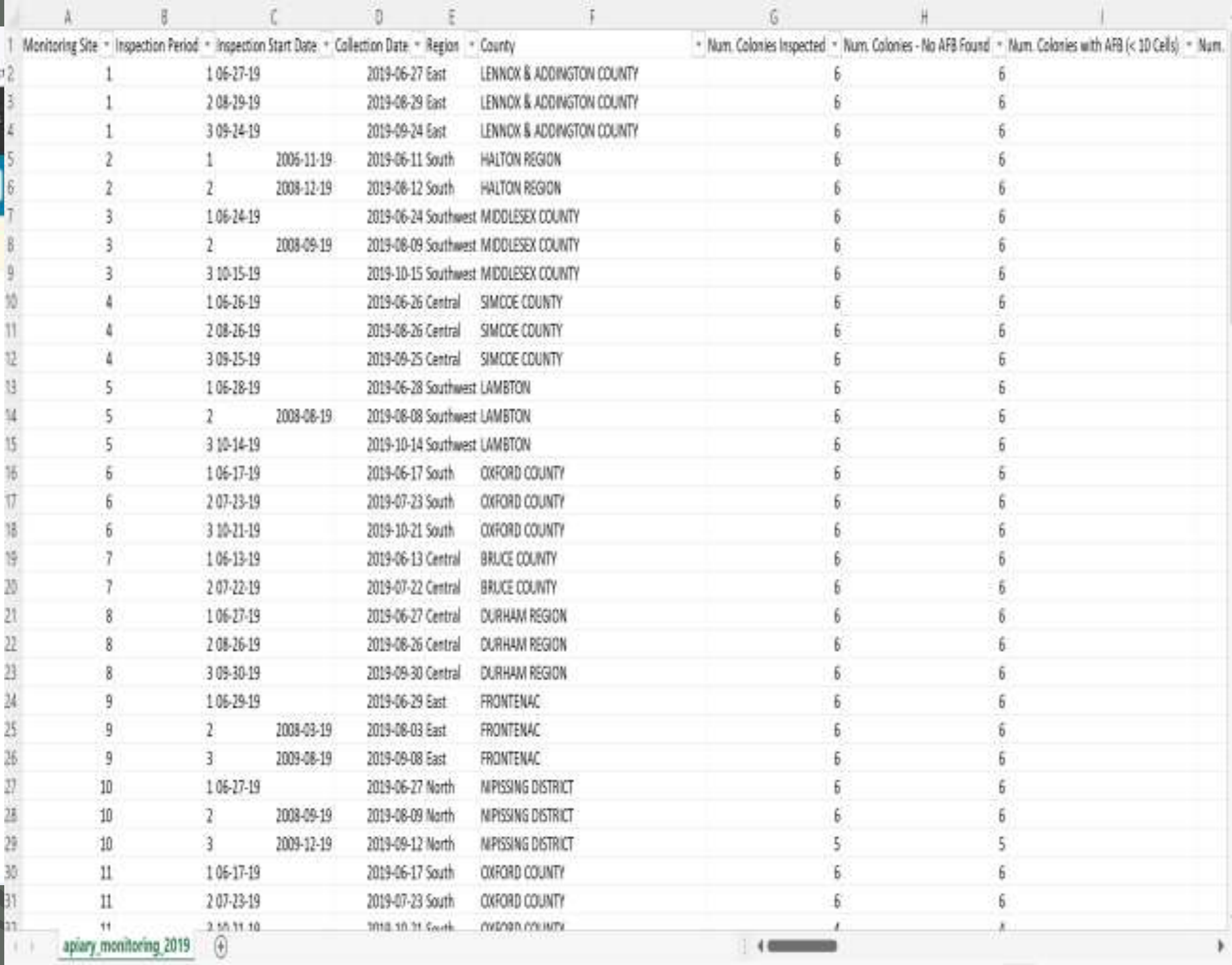
For more information

Contact Agriculture, Food and Rural Affairs

Data

Data Available

The data described here is available for you to use. [Learn more]



	A	B	C	D	E	F	G	H	I	
1	Monitoring Site	Inspection Period	Inspection Start Date	Collection Date	Region	County	Num. Colonies Inspected	Num. Colonies - No AFB Found	Num. Colonies with AFB (< 10 Cells)	Num.
2	1	1	06-27-19	2019-06-27	East	LENNOX & ADDINGTON COUNTY	6	6		
3	1	2	08-29-19	2019-08-29	East	LENNOX & ADDINGTON COUNTY	6	6		
4	1	3	09-24-19	2019-09-24	East	LENNOX & ADDINGTON COUNTY	6	6		
5	2	1	2006-11-19	2019-06-11	South	HALTON REGION	6	6		
6	2	2	2008-12-19	2019-08-12	South	HALTON REGION	6	6		
7	3	1	06-24-19	2019-06-24	Southwest	MIDDLESEX COUNTY	6	6		
8	3	2	2008-09-19	2019-08-09	Southwest	MIDDLESEX COUNTY	6	6		
9	3	3	10-15-19	2019-10-15	Southwest	MIDDLESEX COUNTY	6	6		
10	4	1	06-26-19	2019-06-26	Central	SIMCOE COUNTY	6	6		
11	4	2	08-26-19	2019-08-26	Central	SIMCOE COUNTY	6	6		
12	4	3	09-25-19	2019-09-25	Central	SIMCOE COUNTY	6	6		
13	5	1	06-28-19	2019-06-28	Southwest	LAMBTON	6	6		
14	5	2	2008-08-19	2019-08-08	Southwest	LAMBTON	6	6		
15	5	3	10-14-19	2019-10-14	Southwest	LAMBTON	6	6		
16	6	1	06-17-19	2019-06-17	South	OXFORD COUNTY	6	6		
17	6	2	07-23-19	2019-07-23	South	OXFORD COUNTY	6	6		
18	6	3	10-21-19	2019-10-21	South	OXFORD COUNTY	6	6		
19	7	1	06-13-19	2019-06-13	Central	BRUCE COUNTY	6	6		
20	7	2	07-22-19	2019-07-22	Central	BRUCE COUNTY	6	6		
21	8	1	06-27-19	2019-06-27	Central	DURHAM REGION	6	6		
22	8	2	08-26-19	2019-08-26	Central	DURHAM REGION	6	6		
23	8	3	09-30-19	2019-09-30	Central	DURHAM REGION	6	6		
24	9	1	06-29-19	2019-06-29	East	FRONTENAC	6	6		
25	9	2	2008-03-19	2019-08-03	East	FRONTENAC	6	6		
26	9	3	2009-08-19	2019-09-08	East	FRONTENAC	6	6		
27	10	1	06-27-19	2019-06-27	North	NIPISSING DISTRICT	6	6		
28	10	2	2008-09-19	2019-08-09	North	NIPISSING DISTRICT	6	6		
29	10	3	2009-12-19	2019-09-12	North	NIPISSING DISTRICT	5	5		
30	11	1	06-17-19	2019-06-17	South	OXFORD COUNTY	6	6		
31	11	2	07-23-19	2019-07-23	South	OXFORD COUNTY	6	6		
32	11	3	10-21-19	2019-10-21	South	OXFORD COUNTY	[illegible]	[illegible]		

2

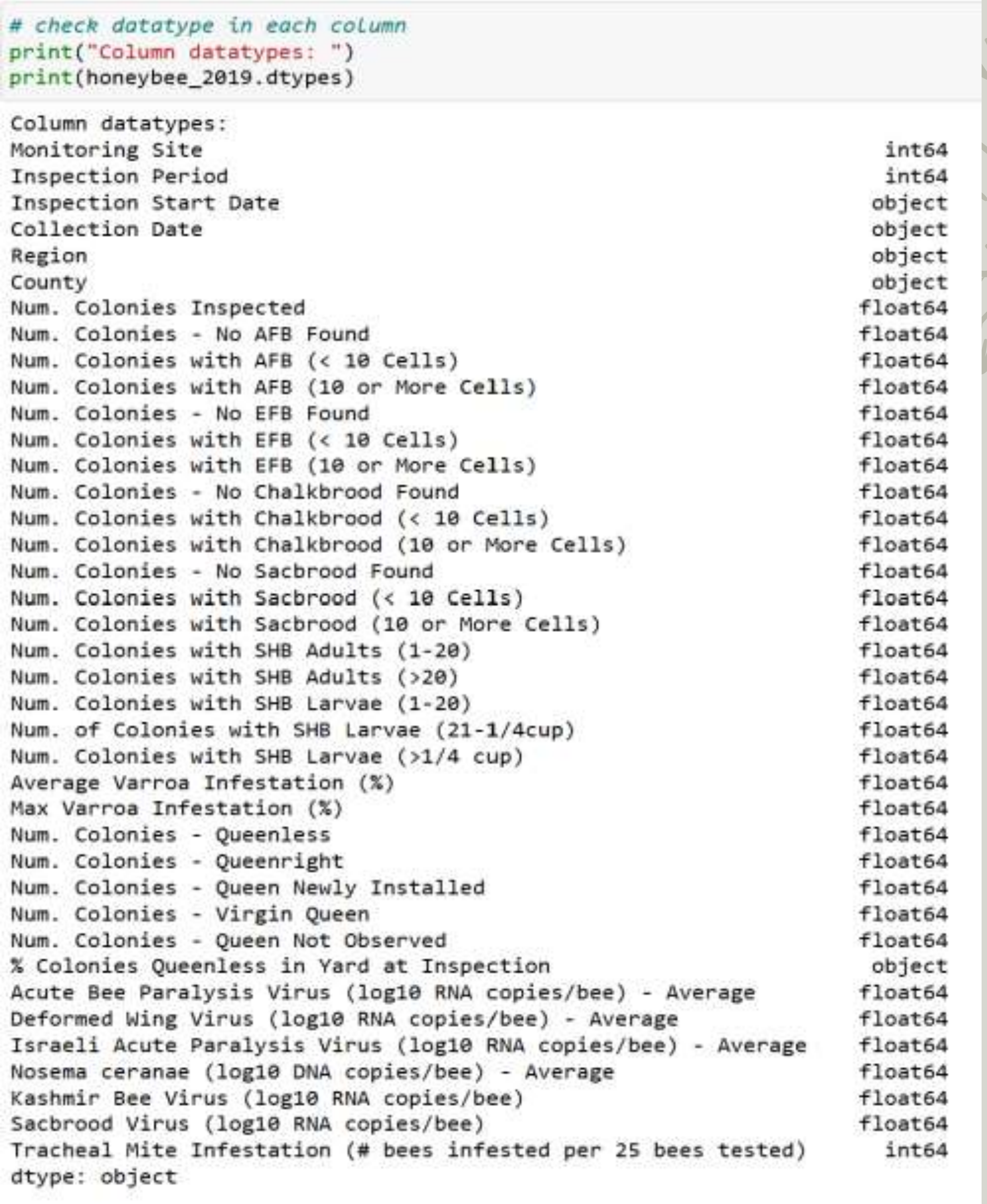
Design

Documentation

1. Project Proposal — Microsoft Word Document
2. Project Charter — Microsoft Word Document
3. Sharepoint Site — https://georgiancollege.sharepoint.com/sites/HoneybeepestsandpathogensinOntario

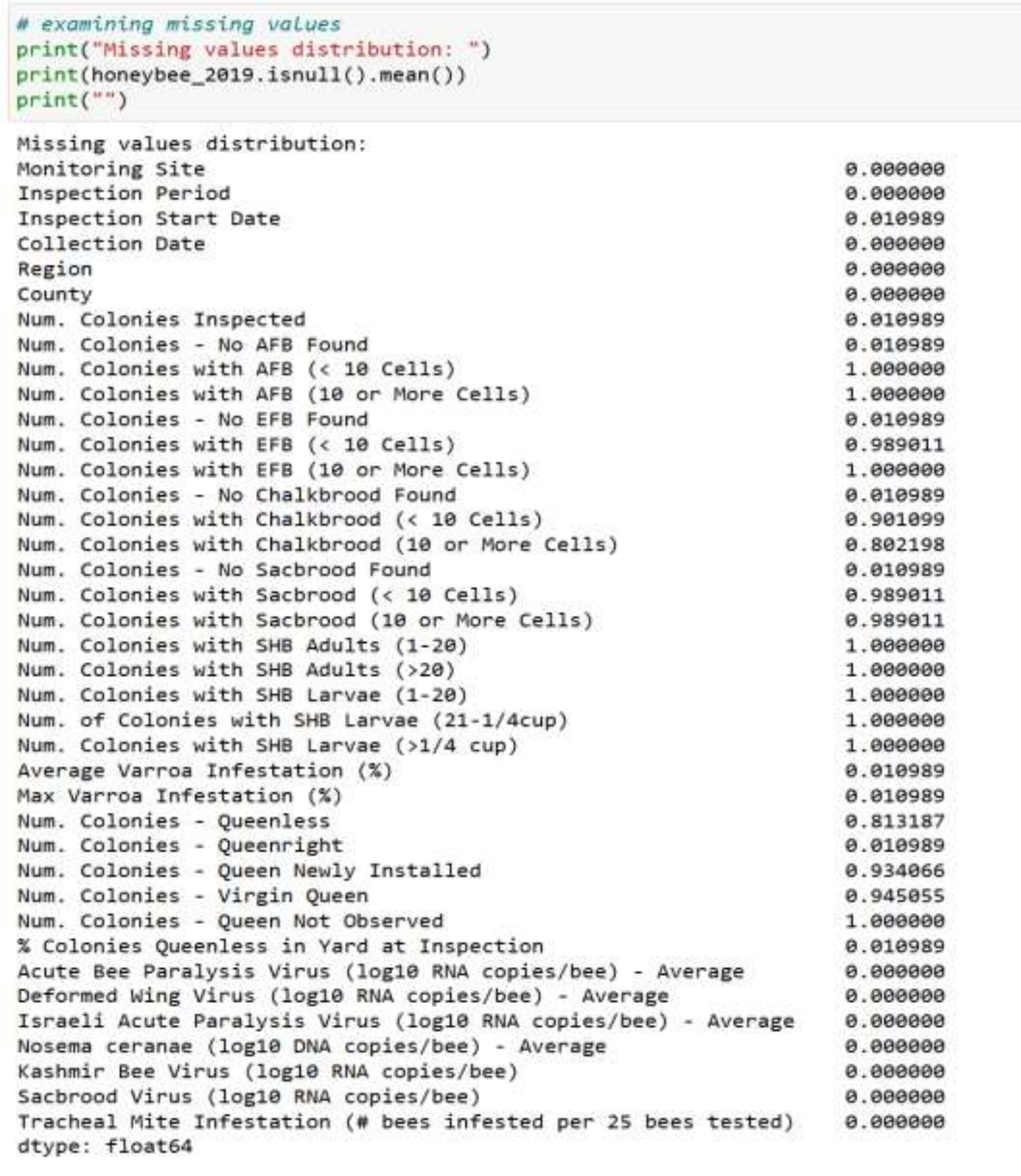
3 Development

Datatypes of variables and missing values distribution for year 2019



```
# check datatype in each column
print("Column datatypes: ")
print(honeybee_2019.dtypes)
```

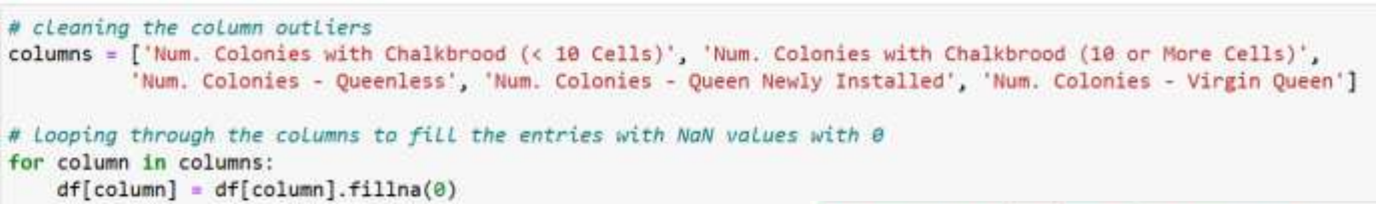
```
Column datatypes:
Monitoring Site                                                     int64
Inspection Period                                                   int64
Inspection Start Date                                              object
Collection Date                                                    object
Region                                                             object
County                                                             object
Num. Colonies Inspected                                           float64
Num. Colonies - No AFB Found                                      float64
Num. Colonies with AFB (< 10 Cells)                               float64
Num. Colonies with AFB (10 or More Cells)                         float64
Num. Colonies - No EFB Found                                      float64
Num. Colonies with EFB (< 10 Cells)                               float64
Num. Colonies with EFB (10 or More Cells)                         float64
Num. Colonies - No Chalkbrood Found                               float64
Num. Colonies with Chalkbrood (< 10 Cells)                        float64
Num. Colonies with Chalkbrood (10 or More Cells)                  float64
Num. Colonies - No Sacbrood Found                                 float64
Num. Colonies with Sacbrood (< 10 Cells)                          float64
Num. Colonies with Sacbrood (10 or More Cells)                    float64
Num. Colonies with SHB Adults (1-20)                              float64
Num. Colonies with SHB Adults (>20)                               float64
Num. Colonies with SHB Larvae (1-20)                              float64
Num. of Colonies with SHB Larvae (21-1/4cup)                      float64
Num. Colonies with SHB Larvae (>1/4 cup)                          float64
Average Varroa Infestation (%)                                    float64
Max Varroa Infestation (%)                                        float64
Num. Colonies - Queenless                                         float64
Num. Colonies - Queenright                                        float64
Num. Colonies - Queen Newly Installed                             float64
Num. Colonies - Virgin Queen                                      float64
Num. Colonies - Queen Not Observed                                float64
% Colonies Queenless in Yard at Inspection                         object
Acute Bee Paralysis Virus (log10 RNA copies/bee) - Average        float64
Deformed Wing Virus (log10 RNA copies/bee) - Average              float64
Israeli Acute Paralysis Virus (log10 RNA copies/bee) - Average    float64
Nosema ceranae (log10 DNA copies/bee) - Average                   float64
Kashmir Bee Virus (log10 RNA copies/bee)                          float64
Sacbrood Virus (log10 RNA copies/bee)                             float64
Tracheal Mite Infestation (# bees infested per 25 bees tested)      int64
dtype: object
```



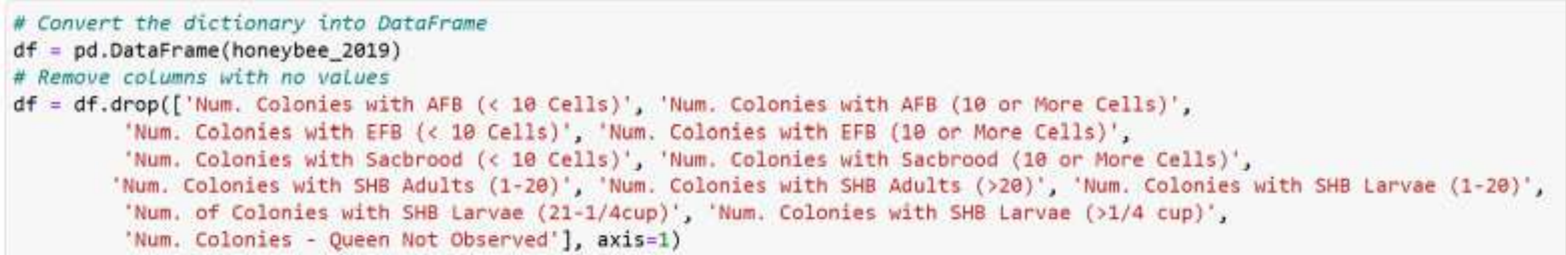
```
# examining missing values
print("Missing values distribution: ")
print(honeybee_2019.isnull().mean())
print("")
```

```
Missing values distribution:
Monitoring Site                                                   0.000000
Inspection Period                                                 0.000000
Inspection Start Date                                             0.010989
Collection Date                                                   0.000000
Region                                                            0.000000
County                                                            0.000000
Num. Colonies Inspected                                           0.010989
Num. Colonies - No AFB Found                                      0.010989
Num. Colonies with AFB (< 10 Cells)                               1.000000
Num. Colonies with AFB (10 or More Cells)                         1.000000
Num. Colonies - No EFB Found                                      0.010989
Num. Colonies with EFB (< 10 Cells)                               0.989011
Num. Colonies with EFB (10 or More Cells)                         1.000000
Num. Colonies - No Chalkbrood Found                               0.010989
Num. Colonies with Chalkbrood (< 10 Cells)                        0.901099
Num. Colonies with Chalkbrood (10 or More Cells)                  0.802198
Num. Colonies - No Sacbrood Found                                 0.010989
Num. Colonies with Sacbrood (< 10 Cells)                          0.989011
Num. Colonies with Sacbrood (10 or More Cells)                    0.989011
Num. Colonies with SHB Adults (1-20)                              1.000000
Num. Colonies with SHB Adults (>20)                               1.000000
Num. Colonies with SHB Larvae (1-20)                              1.000000
Num. of Colonies with SHB Larvae (21-1/4cup)                      1.000000
Num. Colonies with SHB Larvae (>1/4 cup)                          1.000000
Average Varroa Infestation (%)                                    0.010989
Max Varroa Infestation (%)                                        0.010989
Num. Colonies - Queenless                                         0.813187
Num. Colonies - Queenright                                        0.010989
Num. Colonies - Queen Newly Installed                             0.934066
Num. Colonies - Virgin Queen                                      0.945055
Num. Colonies - Queen Not Observed                                1.000000
% Colonies Queenless in Yard at Inspection                        0.010989
Acute Bee Paralysis Virus (log10 RNA copies/bee) - Average        0.000000
Deformed Wing Virus (log10 RNA copies/bee) - Average              0.000000
Israeli Acute Paralysis Virus (log10 RNA copies/bee) - Average    0.000000
Nosema ceranae (log10 DNA copies/bee) - Average                   0.000000
Kashmir Bee Virus (log10 RNA copies/bee)                          0.000000
Sacbrood Virus (log10 RNA copies/bee)                             0.000000
Tracheal Mite Infestation (# bees infested per 25 bees tested)    0.000000
dtype: float64
```

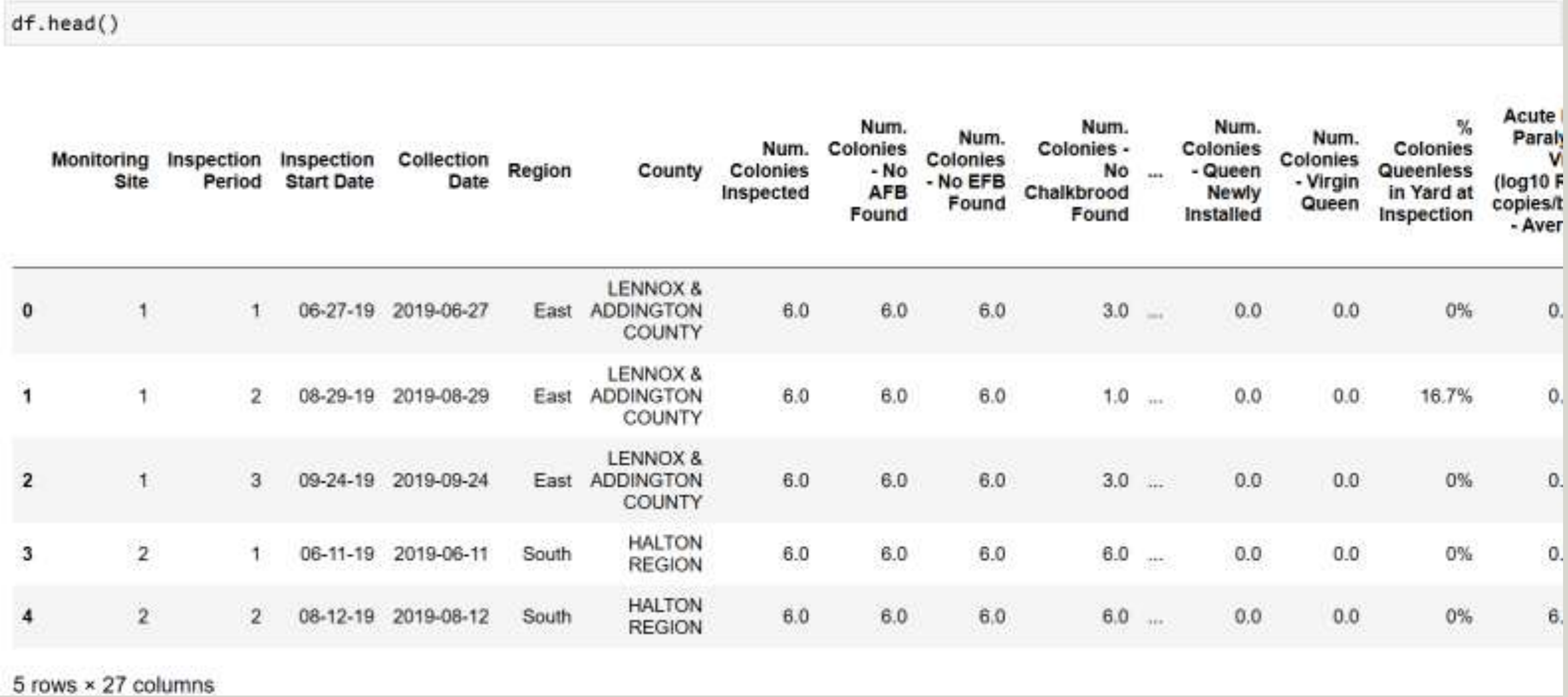
Cleaning Dataset - 2019



```
# cleaning the column outliers
columns = ['Num. Colonies with Chalkbrood (< 10 Cells)', 'Num. Colonies with Chalkbrood (10 or More Cells)',
           'Num. Colonies - Queenless', 'Num. Colonies - Queen Newly Installed', 'Num. Colonies - Virgin Queen']

# looping through the columns to fill the entries with NaN values with 0
for column in columns:
    df[column] = df[column].fillna(0)
```



```
# Convert the dictionary into DataFrame
df = pd.DataFrame(honeybee_2019)
# Remove columns with no values
df = df.drop(['Num. Colonies with AFB (< 10 Cells)', 'Num. Colonies with AFB (10 or More Cells)',
         'Num. Colonies with EFB (< 10 Cells)', 'Num. Colonies with EFB (10 or More Cells)',
         'Num. Colonies with Sacbrood (< 10 Cells)', 'Num. Colonies with Sacbrood (10 or More Cells)',
        'Num. Colonies with SHB Adults (1-20)', 'Num. Colonies with SHB Adults (>20)', 'Num. Colonies with SHB Larvae (1-20)',
         'Num. of Colonies with SHB Larvae (21-1/4cup)', 'Num. Colonies with SHB Larvae (>1/4 cup)',
         'Num. Colonies - Queen Not Observed'], axis=1)
```

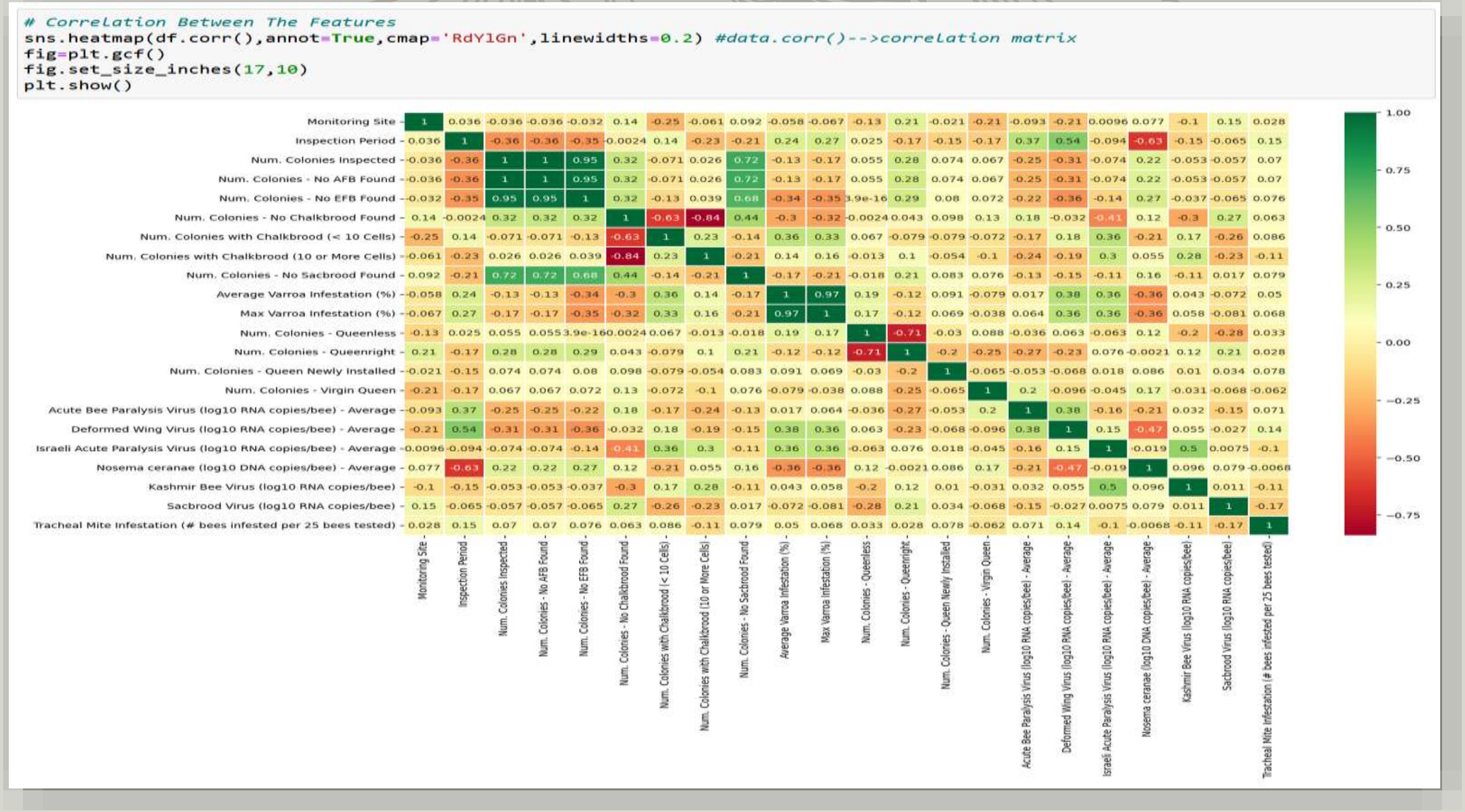


```
df.head()
```

	Monitoring Site	Inspection Period	Inspection Start Date	Collection Date	Region	County	Num. Colonies Inspected	Num. Colonies - No AFB Found	Num. Colonies - No EFB Found	Num. Colonies - No Chalkbrood Found	...	Num. Colonies - Queen Newly Installed	Num. Colonies - Virgin Queen	% Colonies Queenless in Yard at Inspection	Acute Paraly Vi (log10 R copies/t - Aver
0	1	1	06-27-19	2019-06-27	East	LENNOX & ADDINGTON COUNTY	6.0	6.0	6.0	3.0	...	0.0	0.0	0%	0.
1	1	2	08-29-19	2019-08-29	East	LENNOX & ADDINGTON COUNTY	6.0	6.0	6.0	1.0	...	0.0	0.0	16.7%	0.
2	1	3	09-24-19	2019-09-24	East	LENNOX & ADDINGTON COUNTY	6.0	6.0	6.0	3.0	...	0.0	0.0	0%	0.
3	2	1	06-11-19	2019-06-11	South	HALTON REGION	6.0	6.0	6.0	6.0	...	0.0	0.0	0%	0.
4	2	2	08-12-19	2019-08-12	South	HALTON REGION	6.0	6.0	6.0	6.0	...	0.0	0.0	0%	6.

5 rows × 27 columns

Dataset correlation - 2019

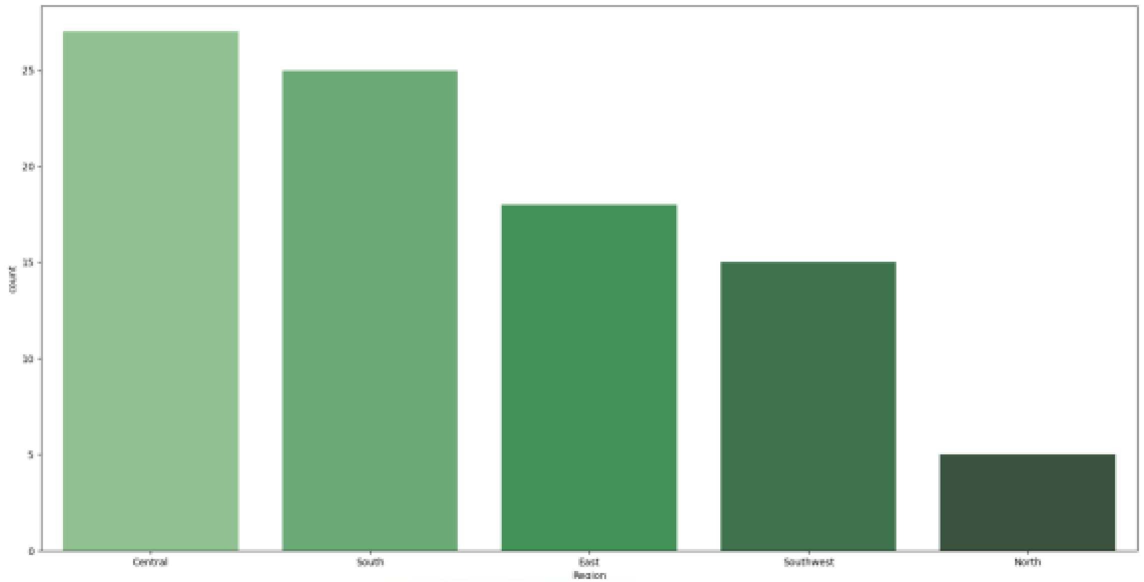


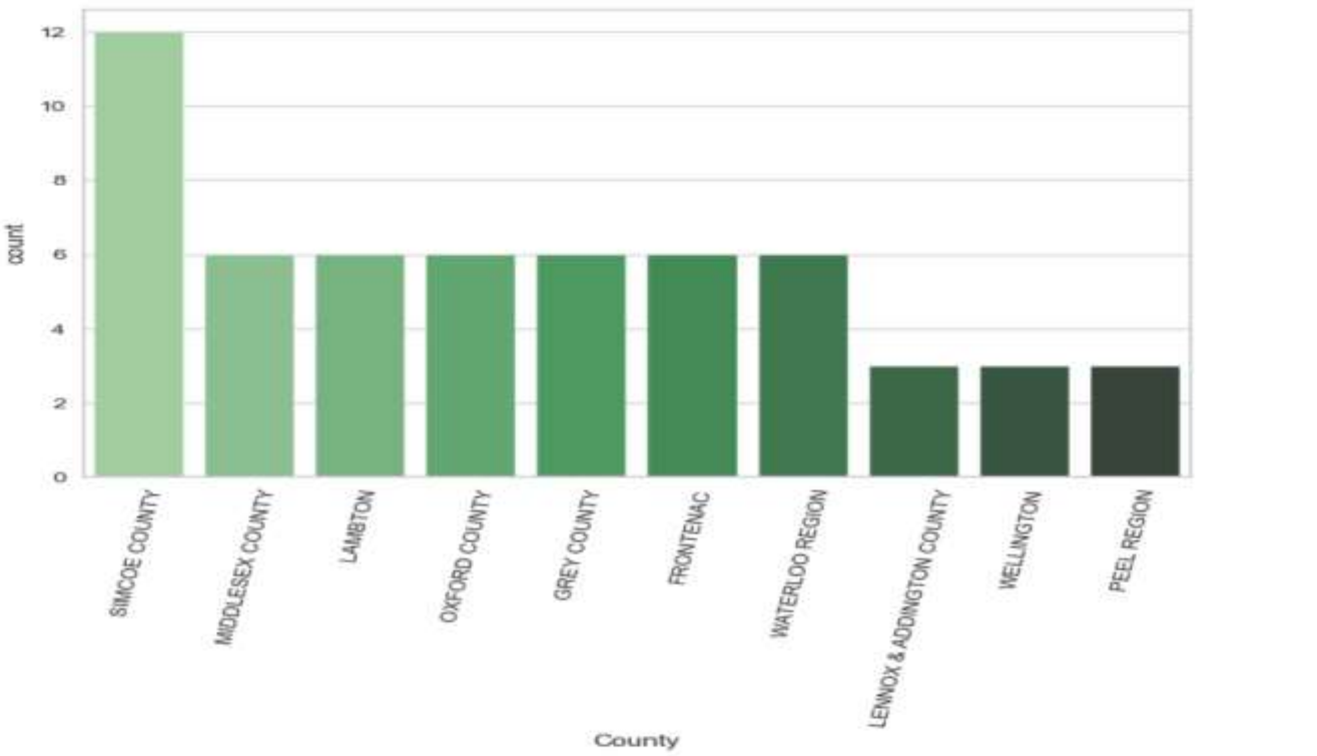
```
# Correlation Between The Features
sns.heatmap(df.corr(),annot=True,cmap='RdYlGn',linewidths=0.2) #data.corr()-->correlation matrix
fig=plt.gcf()
fig.set_size_inches(17,10)
plt.show()
```

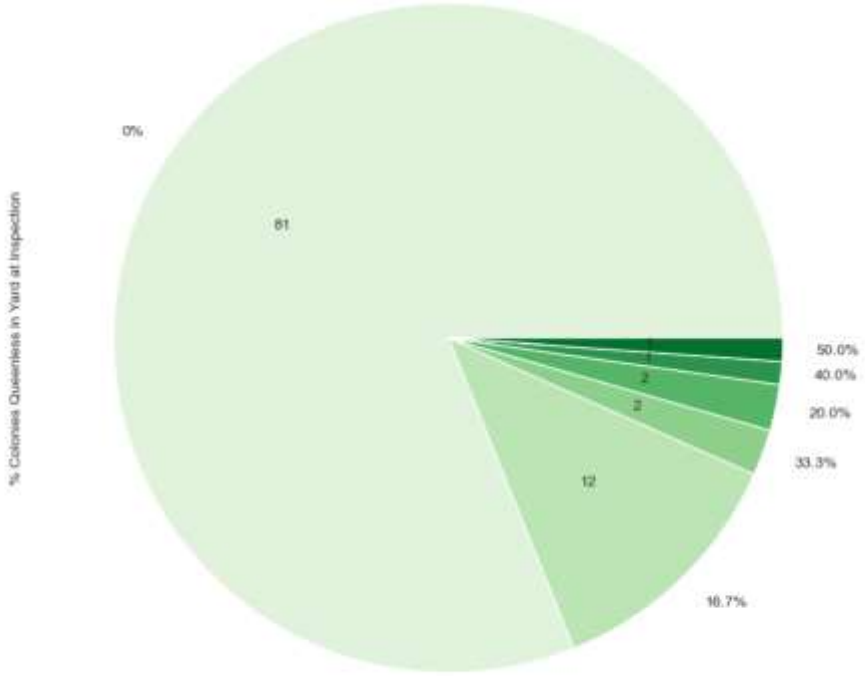
	Monitoring Site	Inspection Period	Num. Colonies Inspected	Num. Colonies - No AFB Found	Num. Colonies - No EFB Found	Num. Colonies - No Chalkbrood Found	Num. Colonies with Chalkbrood (< 10 Cells)	Num. Colonies with Chalkbrood (10 or More Cells)	Num. Colonies - No Sacbrood Found	Average Varroa Infestation (%)	Max Varroa Infestation (%)	Num. Colonies - Queenless	Num. Colonies - Queenright	Num. Colonies - Queen Newly Installed	Num. Colonies - Virgin Queen	Acute Bee Paralysis Virus (log10 RNA copies/bee) - Average	Deformed Wing Virus (log10 RNA copies/bee) - Average	Israeli Acute Paralysis Virus (log10 RNA copies/bee) - Average	Nosema ceranae (log10 DNA copies/bee) - Average	Kashmir Bee Virus (log10 RNA copies/bee)	Sacbrood Virus (log10 RNA copies/bee)	Tracheal Mite Infestation (# bees infested per 25 bees tested)
Monitoring Site	1	0.036	-0.036	-0.036	-0.032	0.14	-0.25	-0.061	0.092	-0.058	-0.067	-0.13	0.21	-0.021	-0.21	-0.093	-0.21	0.0096	0.077	-0.1	0.15	0.028
Inspection Period	0.036	1	-0.36	-0.36	-0.35	-0.0024	0.14	-0.23	-0.21	0.24	0.27	0.025	-0.17	-0.15	-0.17	0.37	0.54	-0.094	-0.63	-0.15	-0.065	0.15
Num. Colonies Inspected	-0.036	-0.36	1	1	0.95	0.32	-0.071	0.026	0.72	-0.13	-0.17	0.055	0.28	0.074	0.067	-0.25	-0.31	-0.074	0.22	-0.053	-0.057	0.07
Num. Colonies - No AFB Found	-0.036	-0.36	1	1	0.95	0.32	-0.071	0.026	0.72	-0.13	-0.17	0.055	0.28	0.074	0.067	-0.25	-0.31	-0.074	0.22	-0.053	-0.057	0.07
Num. Colonies - No EFB Found	-0.032	-0.35	0.95	0.95	1	0.32	-0.13	0.039	0.68	-0.34	-0.35	3.9e-16	0.29	0.08	0.072	-0.22	-0.36	-0.14	0.27	-0.037	-0.065	0.076
Num. Colonies - No Chalkbrood Found	0.14	-0.0024	0.32	0.32	0.32	1	-0.63	-0.84	0.44	-0.3	-0.32	-0.0024	0.043	0.098	0.13	0.18	-0.032	-0.41	0.12	-0.3	0.27	0.063
Num. Colonies with Chalkbrood (< 10 Cells)	-0.25	0.14	-0.071	-0.071	-0.13	-0.63	1	0.23	-0.14	0.36	0.33	0.067	-0.079	-0.079	-0.072	-0.17	0.18	0.36	-0.21	0.17	-0.26	0.086
Num. Colonies with Chalkbrood (10 or More Cells)	-0.061	-0.23	0.026	0.026	0.039	-0.84	0.23	1	-0.21	0.14	0.16	-0.013	0.1	-0.054	-0.1	-0.24	-0.19	0.3	0.055	0.28	-0.23	-0.11
Num. Colonies - No Sacbrood Found	0.092	-0.21	0.72	0.72	0.68	0.44	-0.14	-0.21	1	-0.17	-0.21	-0.018	0.21	0.083	0.076	-0.13	-0.15	-0.11	0.16	-0.11	0.017	0.079
Average Varroa Infestation (%)	-0.058	0.24	-0.13	-0.13	-0.34	-0.3	0.36	0.14	-0.17	1	0.97	0.19	-0.12	0.091	-0.079	0.017	0.38	0.36	-0.36	0.043	-0.072	0.05
Max Varroa Infestation (%)	-0.067	0.27	-0.17	-0.17	-0.35	-0.32	0.33	0.16	-0.21	0.97	1	0.17	-0.12	0.069	-0.038	0.064	0.36	0.36	-0.36	0.058	-0.081	0.068
Num. Colonies - Queenless	-0.13	0.025	0.055	0.055	3.9e-16	0.0024	0.067	-0.013	-0.018	0.19	0.17	1	-0.71	-0.03	0.088	-0.036	0.063	-0.063	0.12	-0.2	-0.28	0.033
Num. Colonies - Queenright	0.21	-0.17	0.28	0.28	0.29	0.043	-0.079	0.1	0.21	-0.12	-0.12	-0.71	1	-0.2	-0.25	-0.27	-0.23	0.076	-0.0021	0.12	0.21	0.028
Num. Colonies - Queen Newly Installed	-0.021	-0.15	0.074	0.074	0.08	0.098	-0.079	-0.054	0.083	0.091	0.069	-0.03	-0.2	1	-0.065	-0.053	-0.068	0.018	0.086	0.01	0.034	0.078
Num. Colonies - Virgin Queen	-0.21	-0.17	0.067	0.067	0.072	0.13	-0.072	-0.1	0.076	-0.079	-0.038	0.088	-0.25	-0.065	1	0.2	-0.096	-0.045	0.17	-0.031	-0.068	-0.062
Acute Bee Paralysis Virus (log10 RNA copies/bee) - Average	-0.093	0.37	-0.25	-0.25	-0.22	0.18	-0.17	-0.24	-0.13	0.017	0.064	-0.036	-0.27	-0.053	0.2	1	0.38	-0.16	-0.21	0.032	-0.15	0.071
Deformed Wing Virus (log10 RNA copies/bee) - Average	-0.21	0.54	-0.31	-0.31	-0.36	-0.032	0.18	-0.19	-0.15	0.38	0.36	0.063	-0.23	-0.068	-0.096	0.38	1	0.15	-0.47	0.055	-0.027	0.14
Israeli Acute Paralysis Virus (log10 RNA copies/bee) - Average	0.0096	-0.094	-0.074	-0.074	-0.14	-0.41	0.36	0.3	-0.11	0.36	0.36	-0.063	0.076	0.018	-0.045	-0.16	0.15	1	-0.019	0.5	0.0075	-0.1
Nosema ceranae (log10 DNA copies/bee) - Average	0.077	-0.63	0.22	0.22	0.27	0.12	-0.21	0.055	0.16	-0.36	-0.36	0.12	-0.0021	0.086	0.17	-0.21	-0.47	-0.019	1	0.096	0.079	-0.0068
Kashmir Bee Virus (log10 RNA copies/bee)	-0.1	-0.15	-0.053	-0.053	-0.037	-0.3	0.17	0.28	-0.11	0.043	0.058	-0.2	0.12	0.01	-0.031	0.032	0.055	0.5	0.096	1	0.011	-0.11
Sacbrood Virus (log10 RNA copies/bee)	0.15	-0.065	-0.057	-0.057	-0.065	0.27	-0.26	-0.23	0.017	-0.072	-0.081	-0.28	0.21	0.034	-0.068	-0.15	-0.027	0.0075	0.079	0.011	1	-0.17
Tracheal Mite Infestation (# bees infested per 25 bees tested)	0.028	0.15	0.07	0.07	0.076	0.063	0.086	-0.11	0.079	0.05	0.068	0.033	0.028	0.078	-0.062	0.071	0.14	-0.1	-0.0068	-0.11	-0.17	1

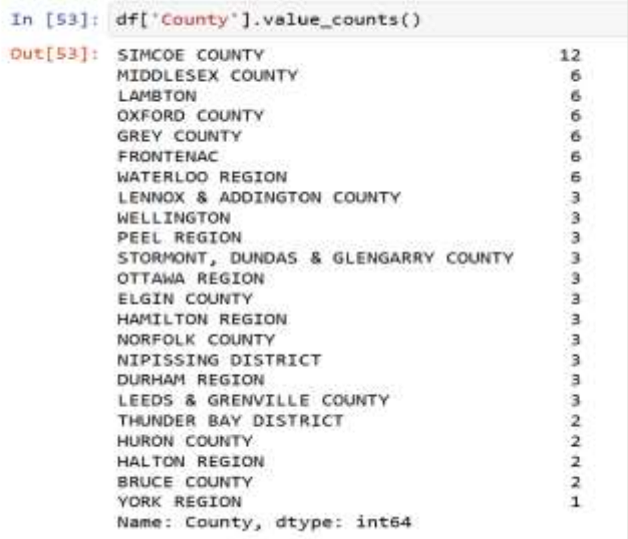
Dataset visualization - 2019

```
<AxesSubplot:xlabel='Region', ylabel='count'>
```



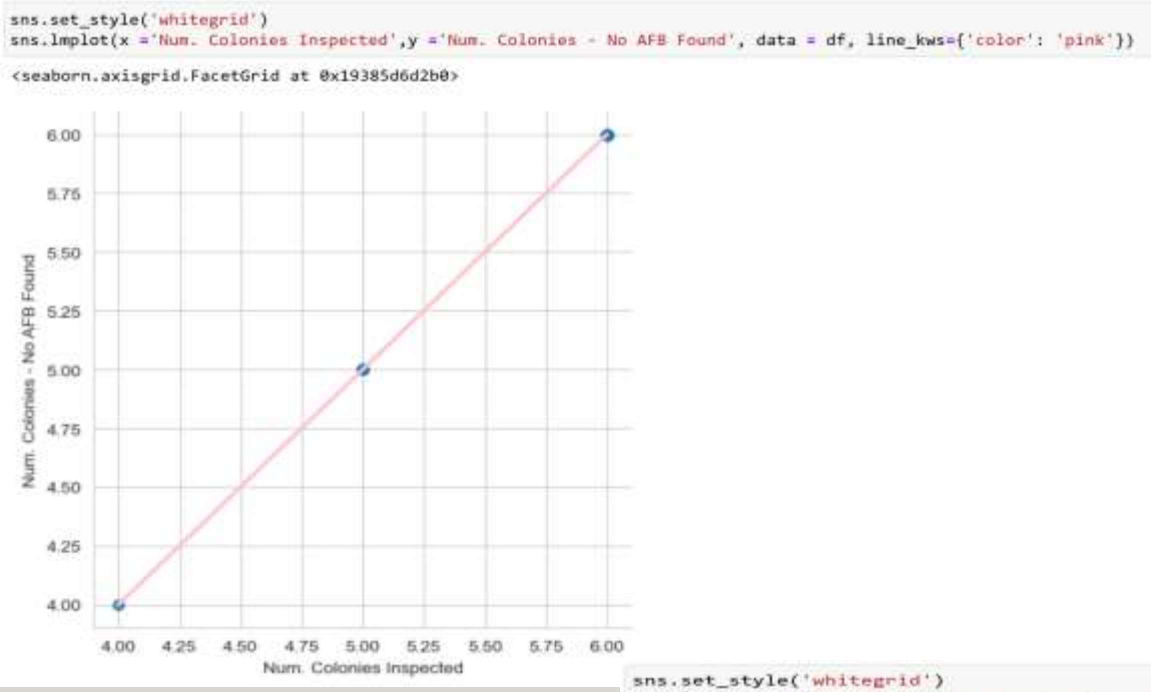






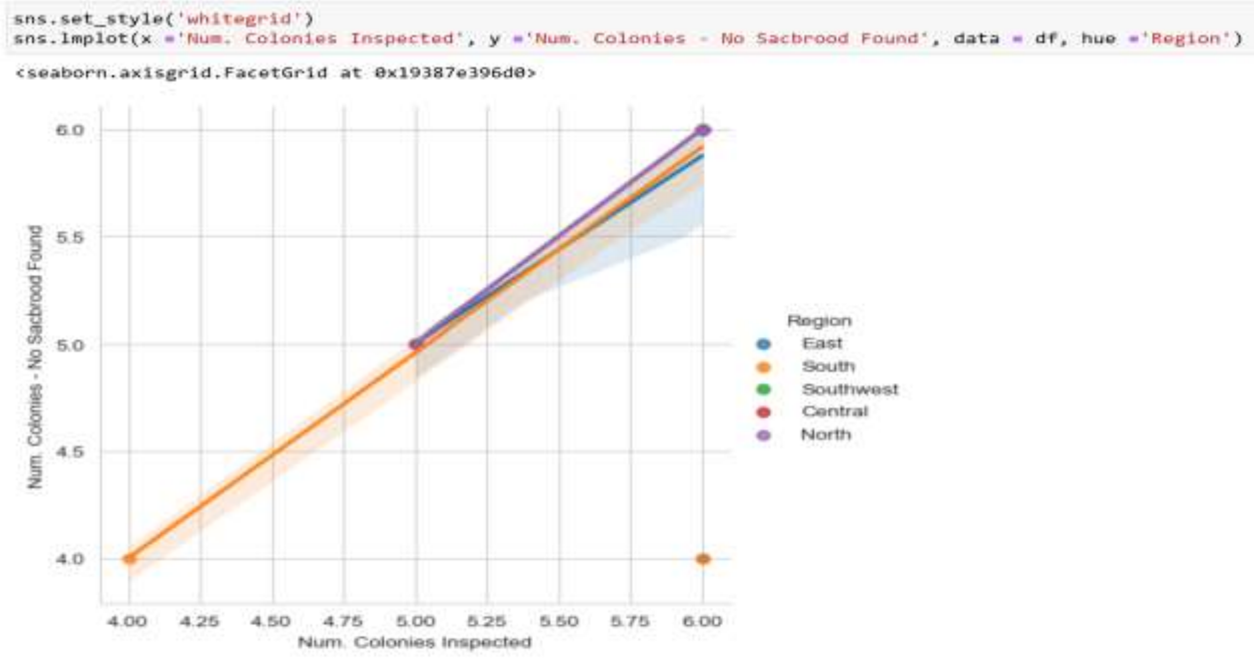
```
In [53]: df['County'].value_counts()

Out[53]: SIMCOE COUNTY                              12
         MIDDLESEX COUNTY                            6
         LAMBTON                                     6
         OXFORD COUNTY                               6
         GREY COUNTY                                 6
         FRONTENAC                                   6
         WATERLOO REGION                             6
         LENNOX & ADDINGTON COUNTY                   3
         WELLINGTON                                  3
         PEEL REGION                                 3
         STORMONT, DUNDAS & GLENGARRY COUNTY         3
         OTTAWA REGION                               3
         ELGIN COUNTY                                3
         HAMILTON REGION                             3
         NORFOLK COUNTY                              3
         NIPISSING DISTRICT                          3
         DURHAM REGION                               3
         LEEDS & GRENVILLE COUNTY                    3
         THUNDER BAY DISTRICT                        2
         HURON COUNTY                                2
         HALTON REGION                               2
         BRUCE COUNTY                                2
         YORK REGION                                 1
         Name: County, dtype: int64
```

Dataset visualization - 2019



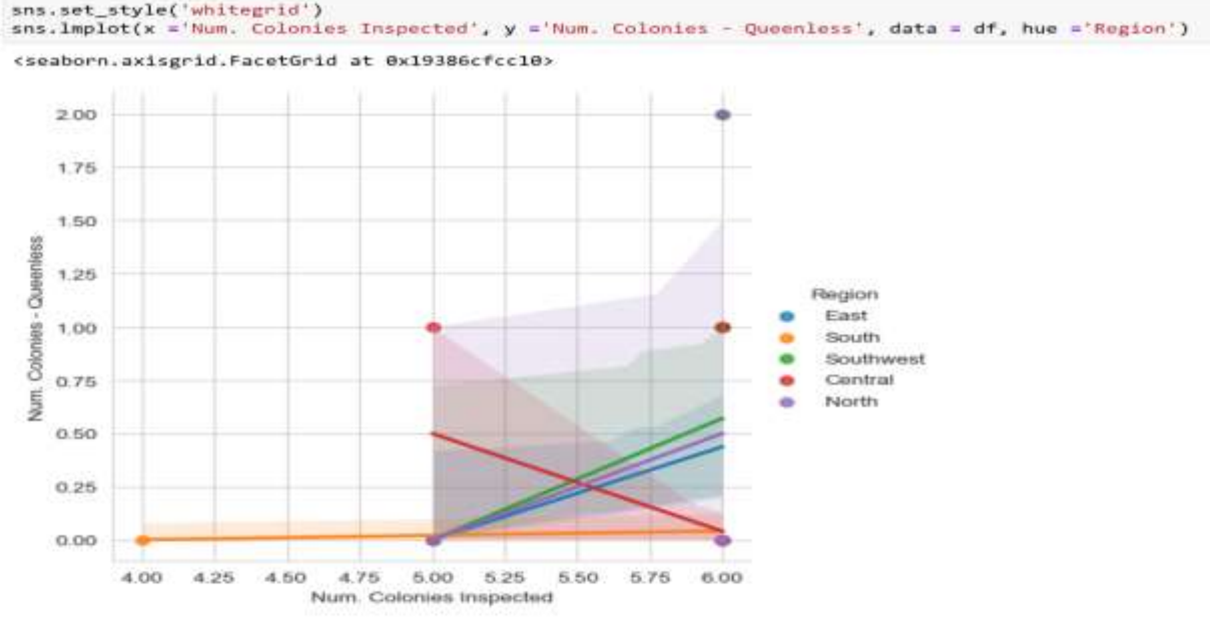
```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected',y ='Num. Colonies - No AFB Found', data = df, line_kws={'color': 'pink'})
```

```
<seaborn.axisgrid.FacetGrid at 0x19385d6d2b0>
```



```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected', y ='Num. Colonies - No Sacbrood Found', data = df, hue ='Region')
```

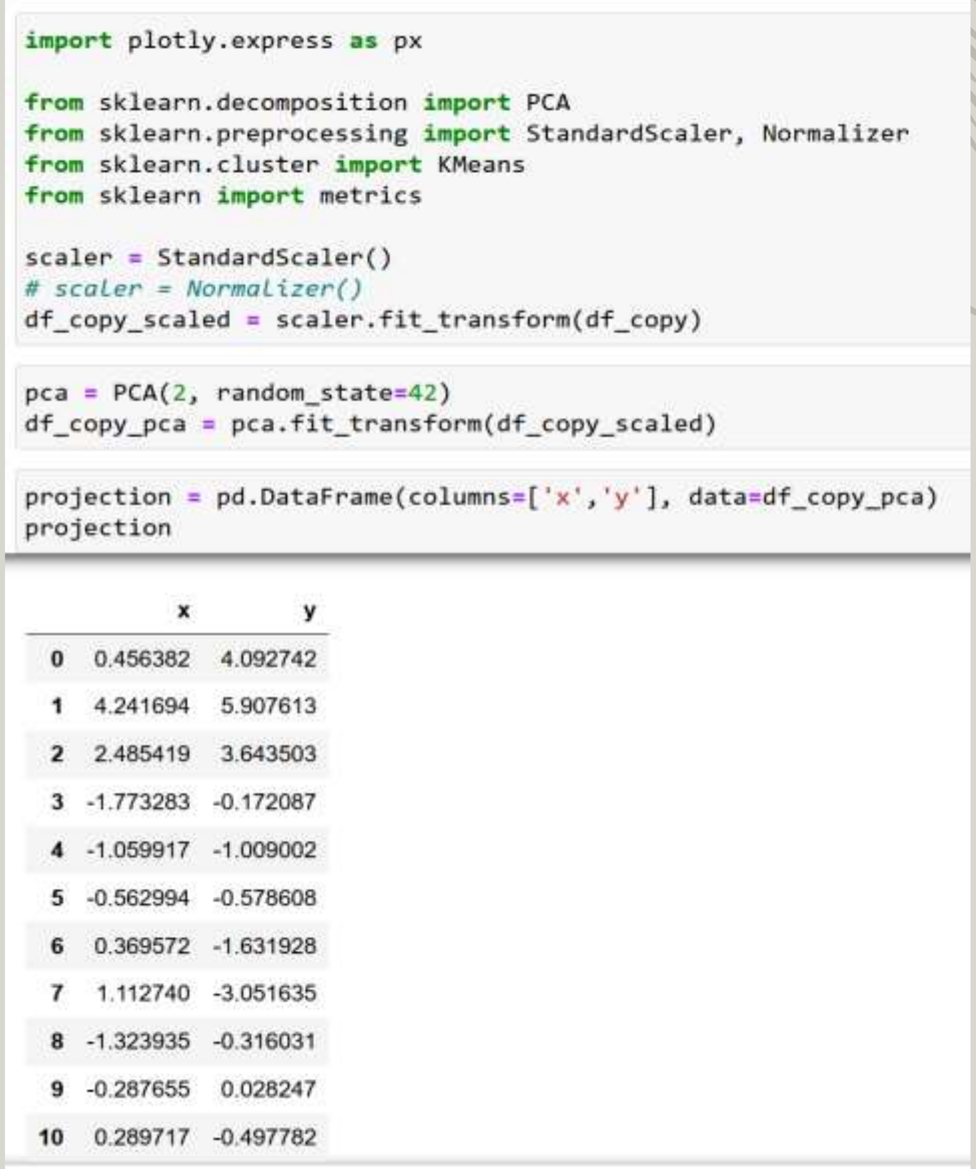
```
<seaborn.axisgrid.FacetGrid at 0x19387e396d0>
```



```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected', y ='Num. Colonies - Queenless', data = df, hue ='Region')
```

```
<seaborn.axisgrid.FacetGrid at 0x19386cfcc10>
```

Predictive analysis - K means clustering

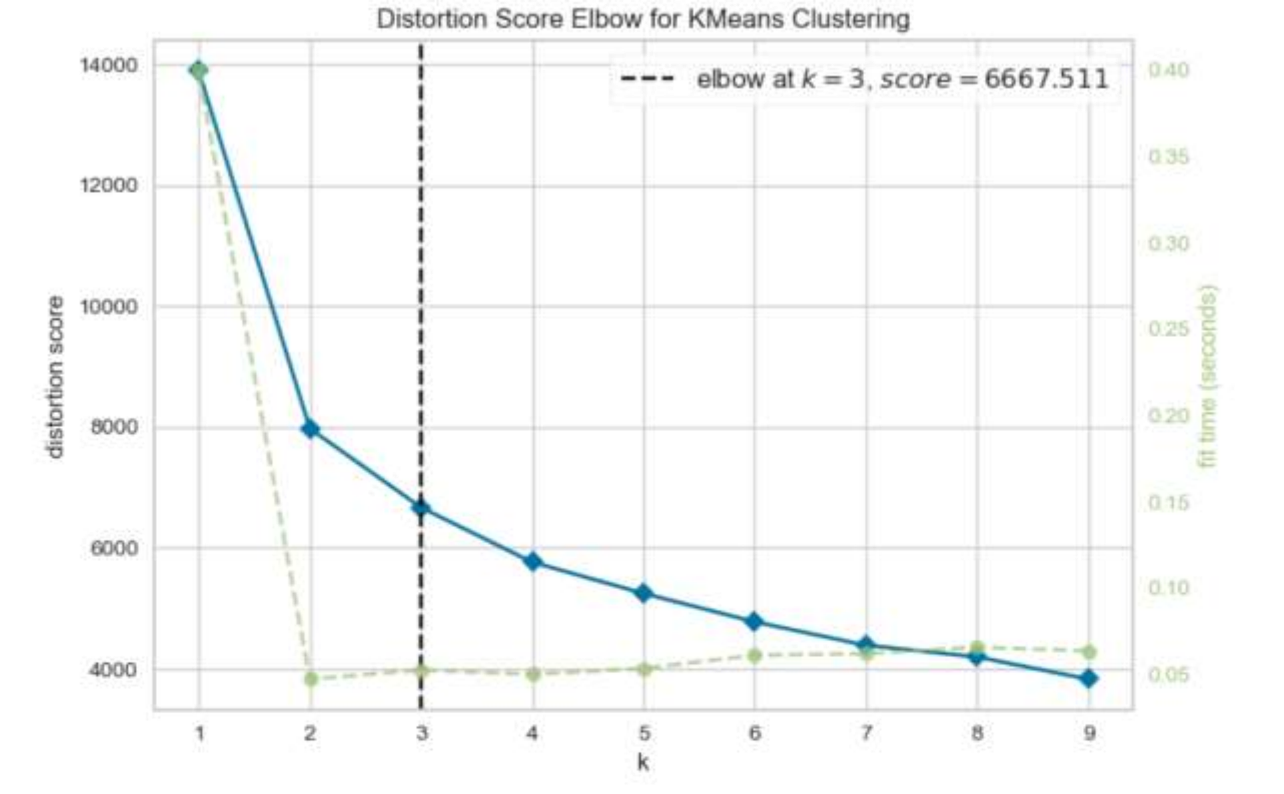


```
import plotly.express as px

from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler, Normalizer
from sklearn.cluster import KMeans
from sklearn import metrics

scaler = StandardScaler()
# scaler = Normalizer()
df_copy_scaled = scaler.fit_transform(df_copy)
```

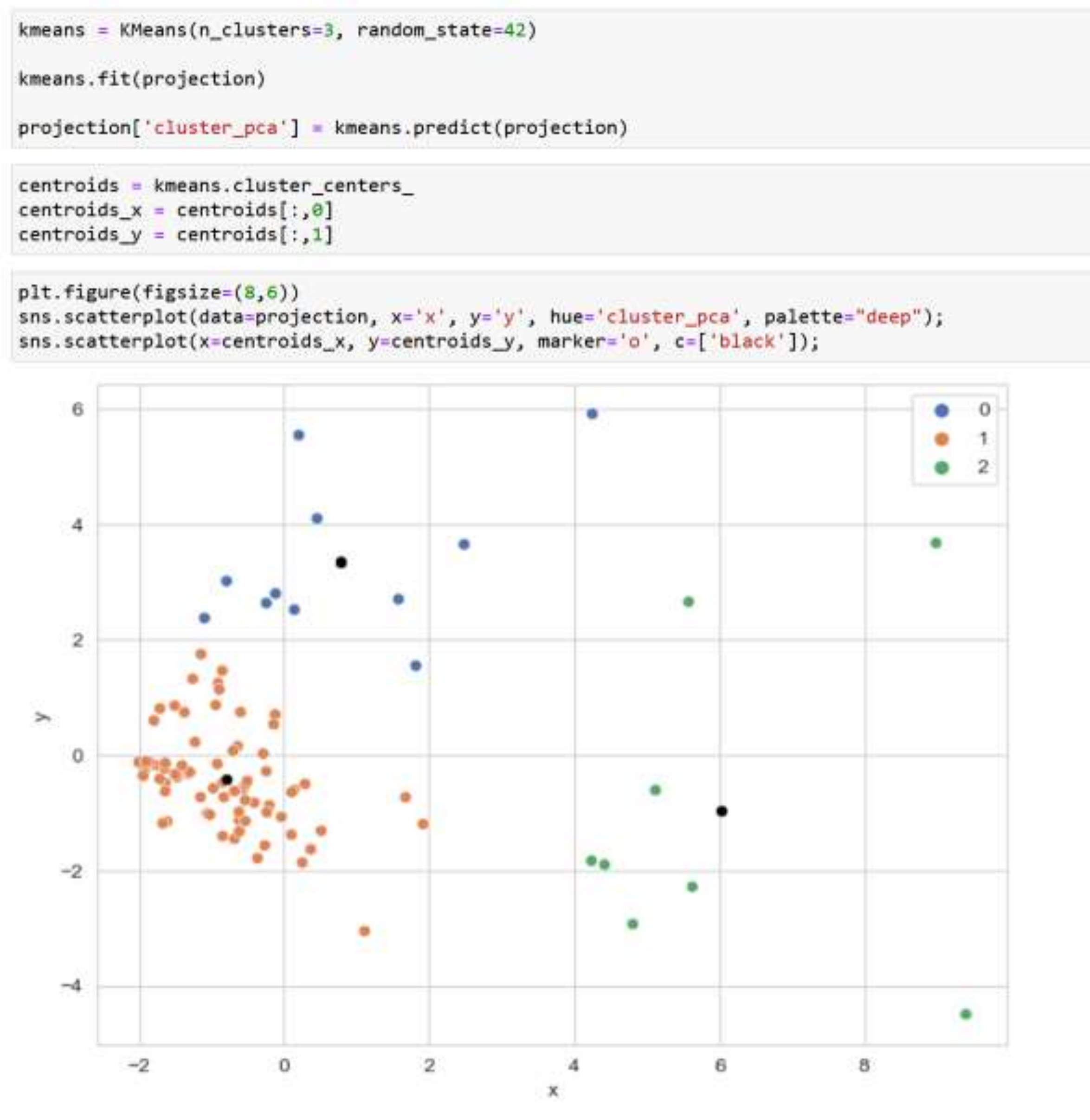
```
pca = PCA(2, random_state=42)
df_copy_pca = pca.fit_transform(df_copy_scaled)
```

```
projection = pd.DataFrame(columns=['x','y'], data=df_copy_pca)
projection
```

	x	y
0	0.456382	4.092742
1	4.241694	5.907613
2	2.485419	3.643503
3	-1.773283	-0.172087
4	-1.059917	-1.009002
5	-0.562994	-0.578608
6	0.369572	-1.631928
7	1.112740	-3.051635
8	-1.323935	-0.316031
9	-0.287655	0.028247
10	0.289717	-0.497782



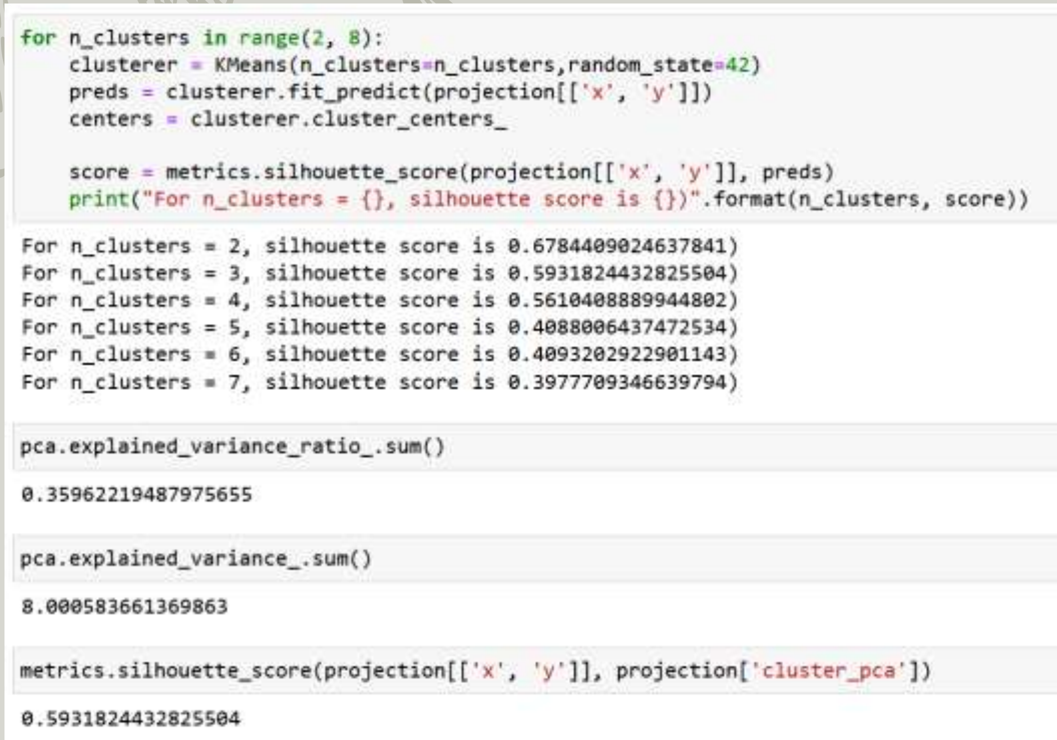
Predictive analysis - K means clustering



```
kmeans = KMeans(n_clusters=3, random_state=42)

kmeans.fit(projection)

projection['cluster_pca'] = kmeans.predict(projection)
```

```
centroids = kmeans.cluster_centers_
centroids_x = centroids[:,0]
centroids_y = centroids[:,1]
```

```
plt.figure(figsize=(8,6))
sns.scatterplot(data=projection, x='x', y='y', hue='cluster_pca', palette="deep");
sns.scatterplot(x=centroids_x, y=centroids_y, marker='o', c=['black']);
```



```
for n_clusters in range(2, 8):
    clusterer = KMeans(n_clusters=n_clusters,random_state=42)
    preds = clusterer.fit_predict(projection[['x', 'y']])
    centers = clusterer.cluster_centers_

    score = metrics.silhouette_score(projection[['x', 'y']], preds)
    print("For n_clusters = {}, silhouette score is {})".format(n_clusters, score))
```

```
For n_clusters = 2, silhouette score is 0.6784409024637841)
For n_clusters = 3, silhouette score is 0.5931824432825504)
For n_clusters = 4, silhouette score is 0.5610408889944802)
For n_clusters = 5, silhouette score is 0.4088006437472534)
For n_clusters = 6, silhouette score is 0.4093202922901143)
For n_clusters = 7, silhouette score is 0.3977709346639794)
```

```
pca.explained_variance_ratio_.sum()
```

```
0.35962219487975655
```

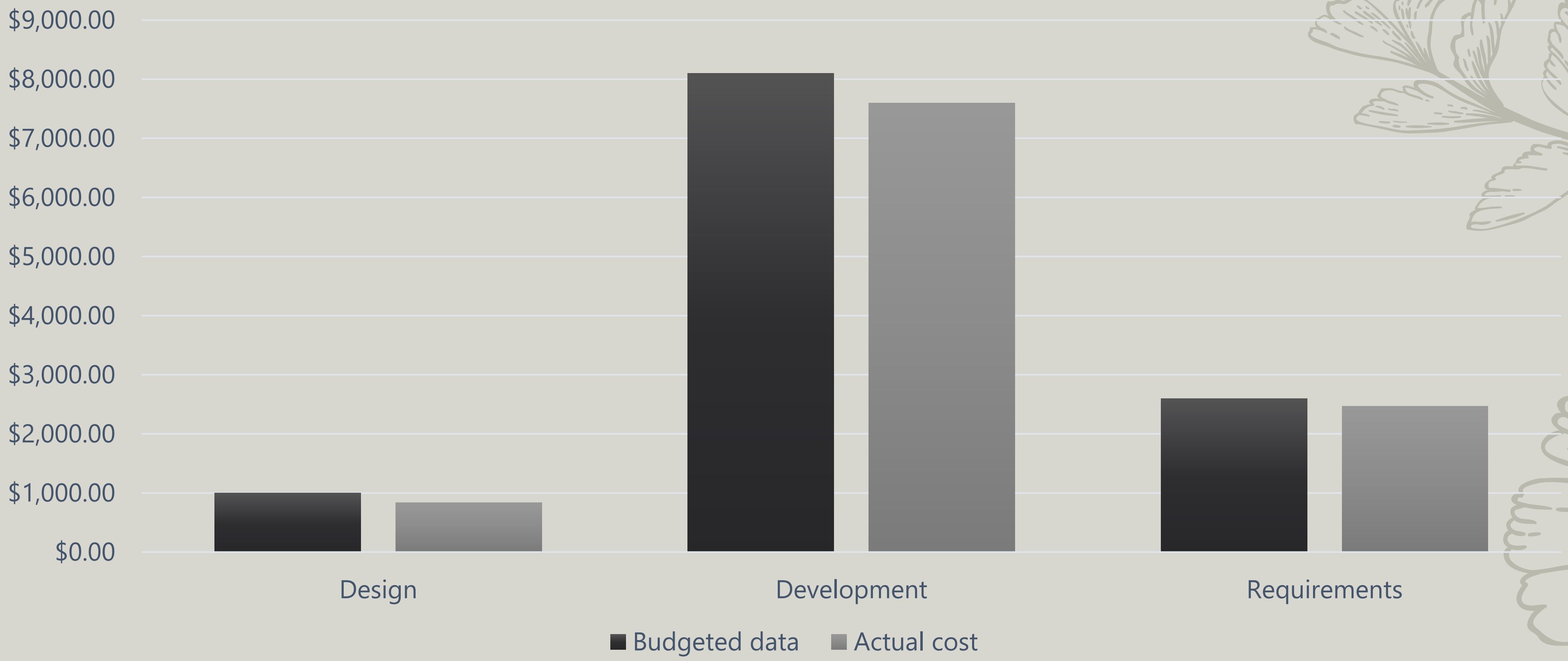
```
pca.explained_variance_.sum()
```

```
8.000583661369863
```

```
metrics.silhouette_score(projection[['x', 'y']], projection['cluster_pca'])
```

```
0.5931824432825504
```

Project Cost



Issues or Challenges encountered this week and what was done to overcome them

We are using Microsoft Excel for cleaning and grouping of data.

Update(25’jan): We are using Python for data cleaning instead of doing manually in Excel.

Understanding outliers and cleaning the data is quite challenging.

Data of years 2017, 2018 and 2019 are considered.

Update(01’feb): no challenges

Update(08’feb): Understanding the numerical data visualization is quite challenging.

Update(15’feb): no challenges

Update(22’feb): Identifying predictive model and testing the accuracy is quite challenging.

Communications

Weekly status meeting with Professor Rick Lambroff

Week – 1 (18’Jan’2023)

- Professor suggested to use Python for cleaning of dataset instead of doing it manually by Microsoft Excel
- Professor provided tutorial sites for ETL of data processing using Python

Week – 2 (25’Jan’2023)

- Professor provided tutorial sites for building a predictive model
- Professor suggested to learn these models and understand clustering algorithms

Week – 3 (01’Feb’2023)

- Professor suggested to add more data visualizations after data cleaning process for a better understanding

Week – 4 (08’Feb’2023)

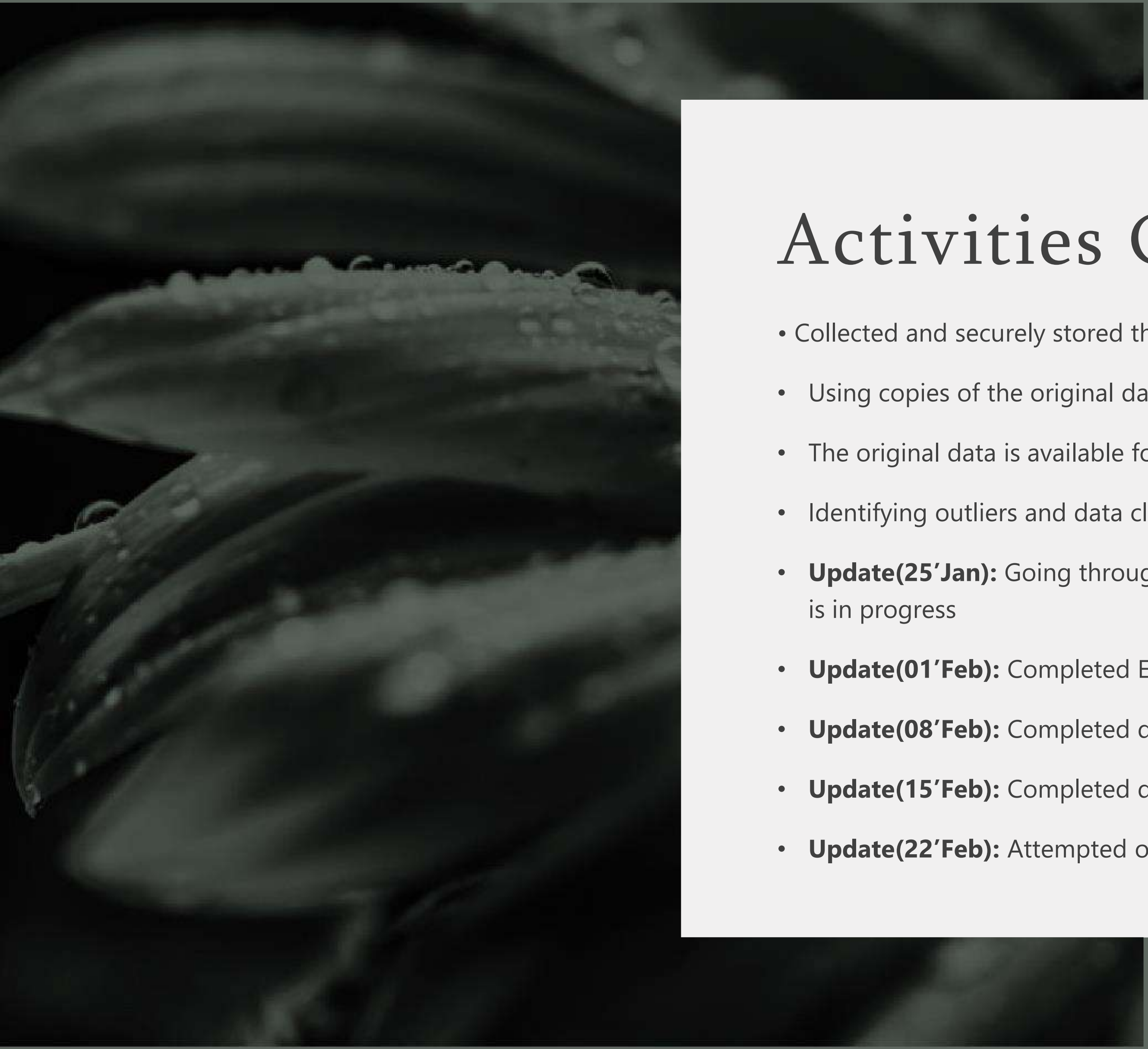
- Professor mentioned few changes in the visualizations like adding heatmap, adding same palette colors

Week – 5 (15’Feb’2023)

- Professor suggested to try one of the predictive models and test for the accuracy

Team meetings

Date	Agenda	Budgeted hours	Attendees	Approval of previous minutes
15/02/2023	Weekly status update – week 5	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
08/02/2023	Weekly status update – week 4	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
01/02/2023	Weekly status update – week 3	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
25/01/2023	Weekly status update – week 2	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
18/01/2023	Weekly status update – week 1	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
07/12/2022	Final group project – submission of SharePoint link, project charter and project proposal	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
23/11/2022	Review of MRP SharePoint Site Follow-up	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
16/11/2022	Review of MRP SharePoint Site Follow-up	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval
09/11/2022	Introductory Client meeting - Finalized project topic and dataset	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Awaiting approval

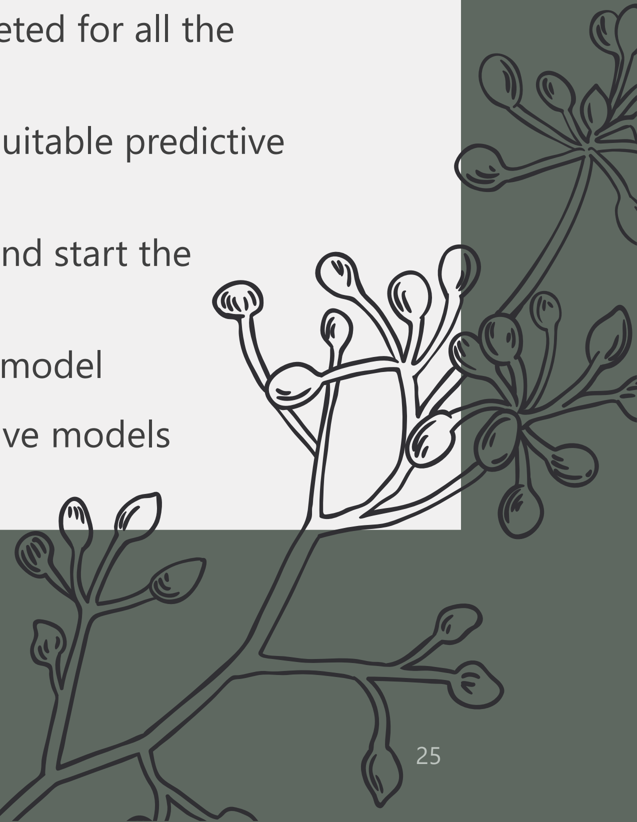


Activities Completed This week

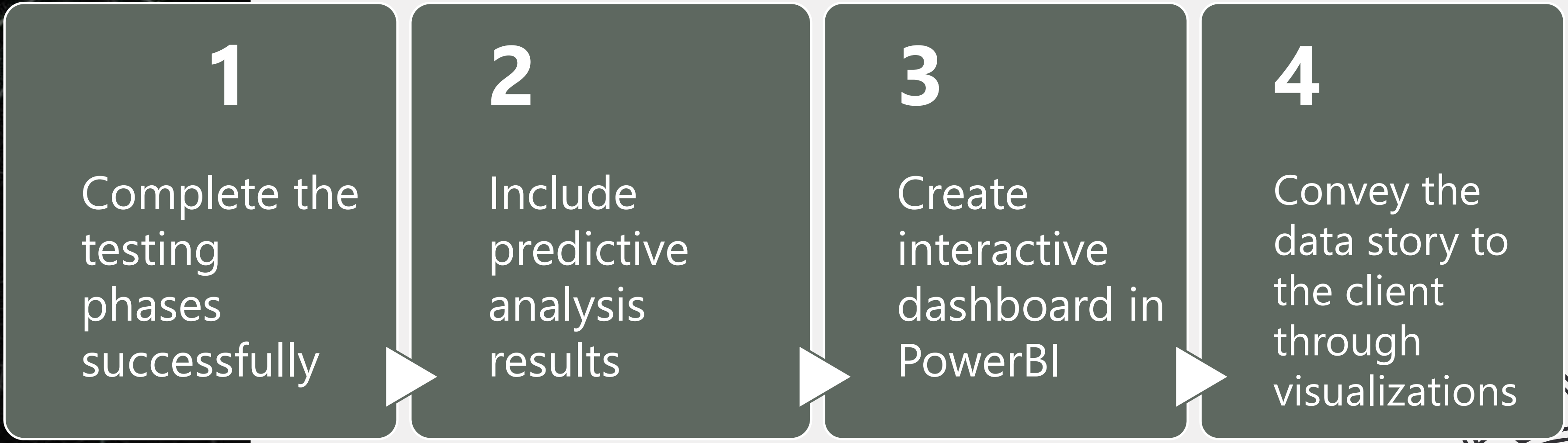
- Collected and securely stored the original data
- Using copies of the original data, clean and prepare the data for analysis
- The original data is available for the years 2017, 2018 and 2019
- Identifying outliers and data cleaning is completed for the year 2017 using Microsoft Excel
- **Update(25′Jan):** Going through tutorials for ETL of data cleaning instead of manual cleaning is in progress
- **Update(01′Feb):** Completed ETL tutorials and data cleaning for the years 2017, 2018, 2019
- **Update(08′Feb):** Completed data visualization for the year 2019
- **Update(15′Feb):** Completed data visualization for the year 2019, 2018, 2017
- **Update(22′Feb):** Attempted one of the predictive models – K means clustering

Activities to be Completed Before Next Report

- Preliminary data analysis is to be completed for all the years 2017, 2018 and 2019
- Securely store the cleaned data using naming conventions and version controls
- Identify the databases, languages to be used and develop a functional flow of the project
- **Update(25’Jan):** Data cleaning using ETL python will be completed for all the datasets of years 2017, 2018 and 2019
- **Update(01’Feb):** Understanding predictive models and find a suitable predictive model for our project
- **Update(08’Feb):** Complete the data visualization for all years and start the development of predictive model
- **Update(15’Feb):** Continue development phase II of prediction model
- **Update(22’Feb):** Continue development and testing of predictive models



Plans for the next phase



1

Complete the testing phases successfully

2

Include predictive analysis results

3

Create interactive dashboard in PowerBI

4

Convey the data story to the client through visualizations

Thank you