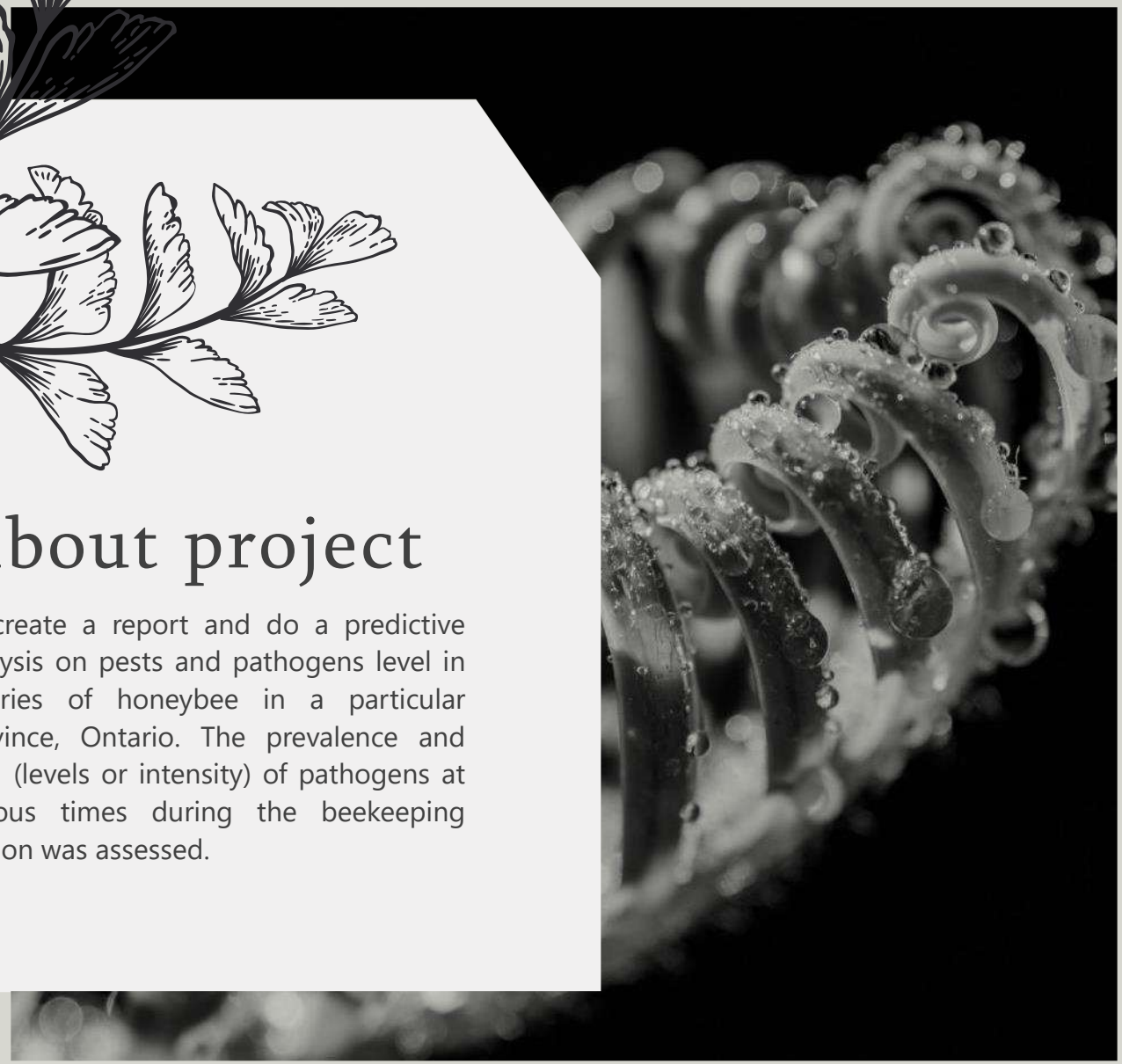
HONEYBEE PESTS & PATHOGENS IN ONTARIO APIARIES

Moganaviniith Rathinavel
Ragavi Mudaliyar
Paras Gangani

Table of Contents



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.

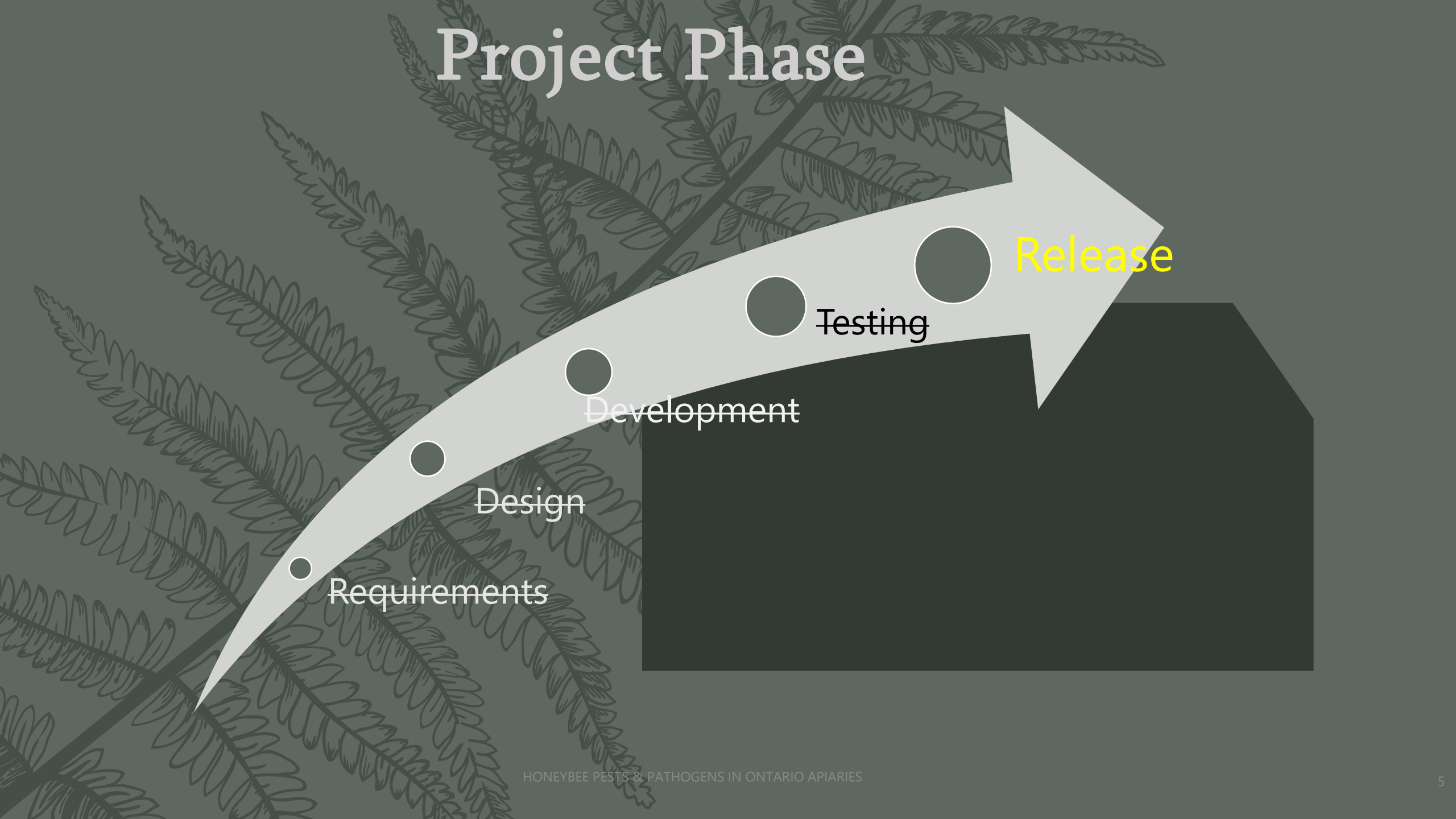
Project Status

Completed

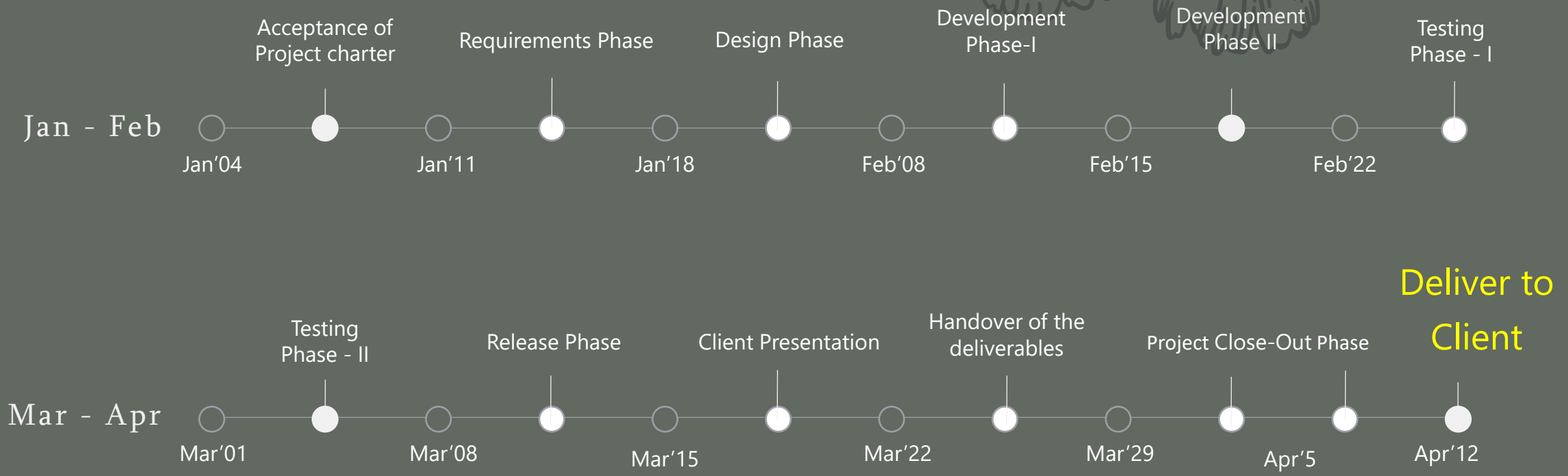
Date of report: 10th April. 2023

Date of last report: 29th March. 2023

Project Phase



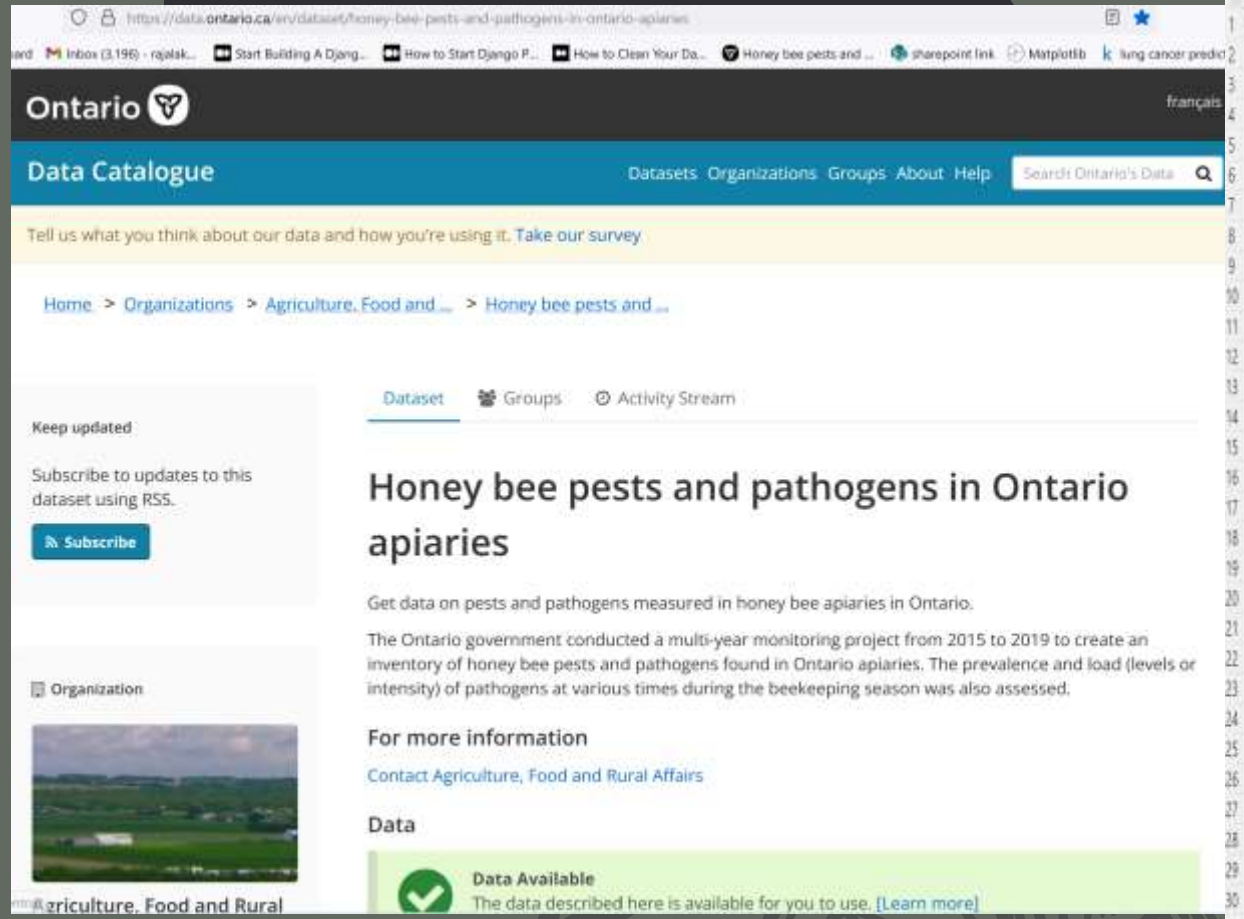
Milestones

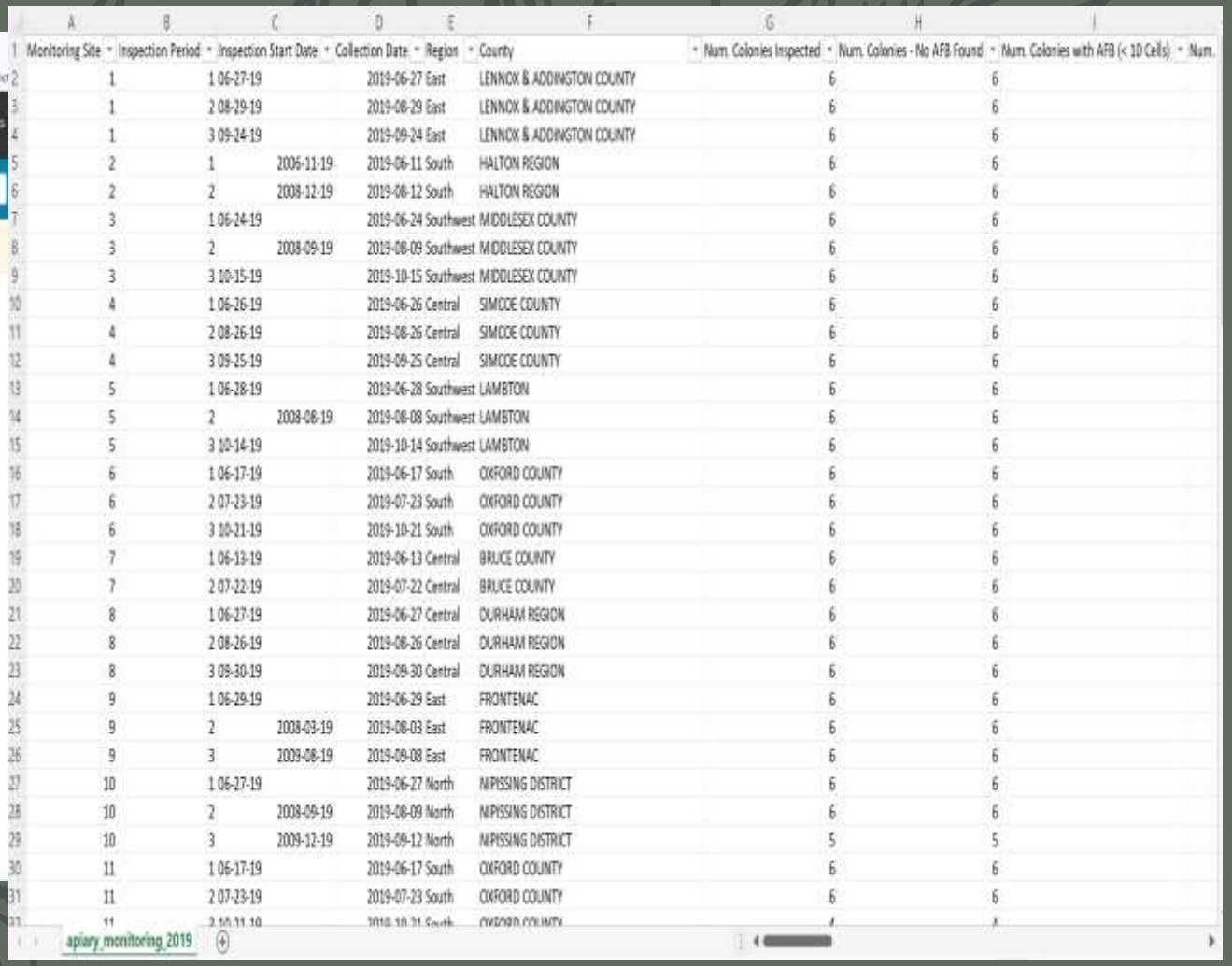

Jan - Feb
Acceptance of Project charter
Requirements Phase
Design Phase
Development Phase-I
Development Phase II
Testing Phase - I
Jan'04
Jan'11
Jan'18
Feb'08
Feb'15
Feb'22
Mar - Apr
Testing Phase - II
Release Phase
Client Presentation
Handover of the deliverables
Project Close-Out Phase
Deliver to Client
Mar'01
Mar'08
Mar'15
Mar'22
Mar'29
Apr'5
Apr'12

1 Requirements

About the Dataset

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





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.
1	1	06-27-19	2019-06-27	East	LENNOX & ADDINGTON COUNTY	6	6		
1	2	08-29-19	2019-08-29	East	LENNOX & ADDINGTON COUNTY	6	6		
1	3	09-24-19	2019-09-24	East	LENNOX & ADDINGTON COUNTY	6	6		
2	1	2006-11-19	2019-06-11	South	HALTON REGION	6	6		
2	2	2008-12-19	2019-08-12	South	HALTON REGION	6	6		
3	1	06-24-19	2019-06-24	Southwest	MIDDLESEX COUNTY	6	6		
3	2	2008-09-19	2019-08-09	Southwest	MIDDLESEX COUNTY	6	6		
3	3	10-15-19	2019-10-15	Southwest	MIDDLESEX COUNTY	6	6		
4	1	06-26-19	2019-06-26	Central	SIMCOE COUNTY	6	6		
4	2	08-26-19	2019-08-26	Central	SIMCOE COUNTY	6	6		
4	3	09-25-19	2019-09-25	Central	SIMCOE COUNTY	6	6		
5	1	06-28-19	2019-06-28	Southwest	LAMBTON	6	6		
5	2	2008-08-19	2019-08-08	Southwest	LAMBTON	6	6		
5	3	10-14-19	2019-10-14	Southwest	LAMBTON	6	6		
6	1	06-17-19	2019-06-17	South	OXFORD COUNTY	6	6		
6	2	07-23-19	2019-07-23	South	OXFORD COUNTY	6	6		
6	3	10-21-19	2019-10-21	South	OXFORD COUNTY	6	6		
7	1	06-13-19	2019-06-13	Central	BRUCE COUNTY	6	6		
7	2	07-22-19	2019-07-22	Central	BRUCE COUNTY	6	6		
8	1	06-27-19	2019-06-27	Central	DURHAM REGION	6	6		
8	2	08-26-19	2019-08-26	Central	DURHAM REGION	6	6		
8	3	09-30-19	2019-09-30	Central	DURHAM REGION	6	6		
9	1	06-29-19	2019-06-29	East	FRONTENAC	6	6		
9	2	2008-03-19	2019-08-03	East	FRONTENAC	6	6		
9	3	2009-08-19	2019-09-08	East	FRONTENAC	6	6		
10	1	06-27-19	2019-06-27	North	NIPISSING DISTRICT	6	6		
10	2	2008-09-19	2019-08-09	North	NIPISSING DISTRICT	6	6		
10	3	2009-12-19	2019-09-12	North	NIPISSING DISTRICT	5	5		
11	1	06-17-19	2019-06-17	South	OXFORD COUNTY	6	6		
11	2	07-23-19	2019-07-23	South	OXFORD COUNTY	6	6		

apiary_monitoring_2019

2 Design

Documentation

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

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()
```

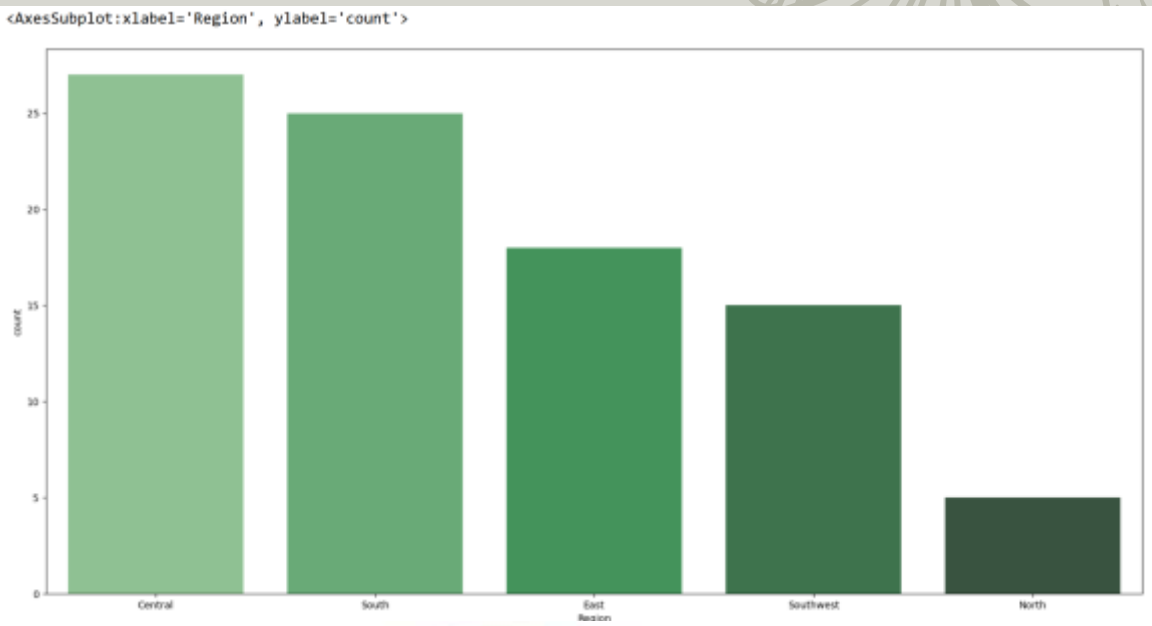
	Monitoring Site	Inspection Period	Inspection Start Date	Collection Date	Region	County	Num. Colonies Inspected	Num. Colonies - No AFB Found	Num. Colonies - No EFB Found	Num. Colonies - No Chalkbrood Found	...	Num. Colonies - Queen Newly Installed	Num. Colonies - Virgin Queen	% Colonies Queenless in Yard at Inspection	Acute Paral V (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.

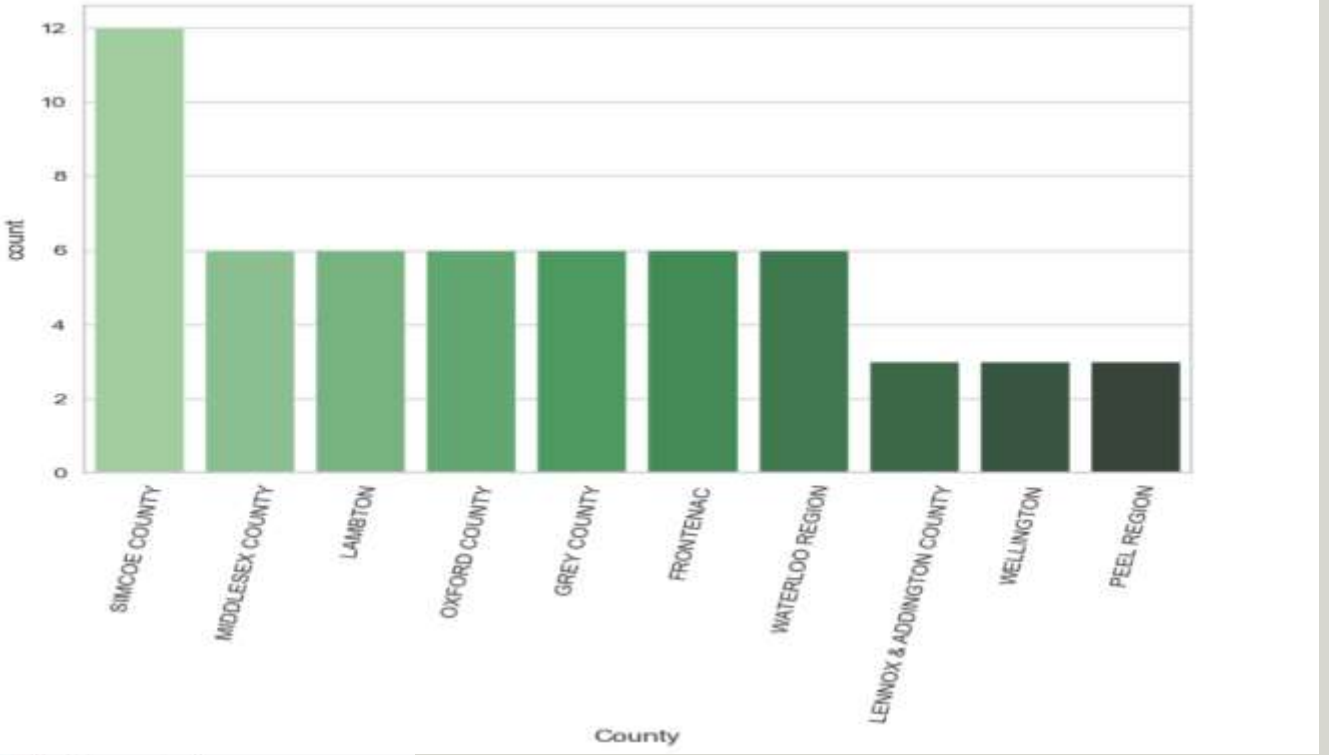
5 rows × 27 columns

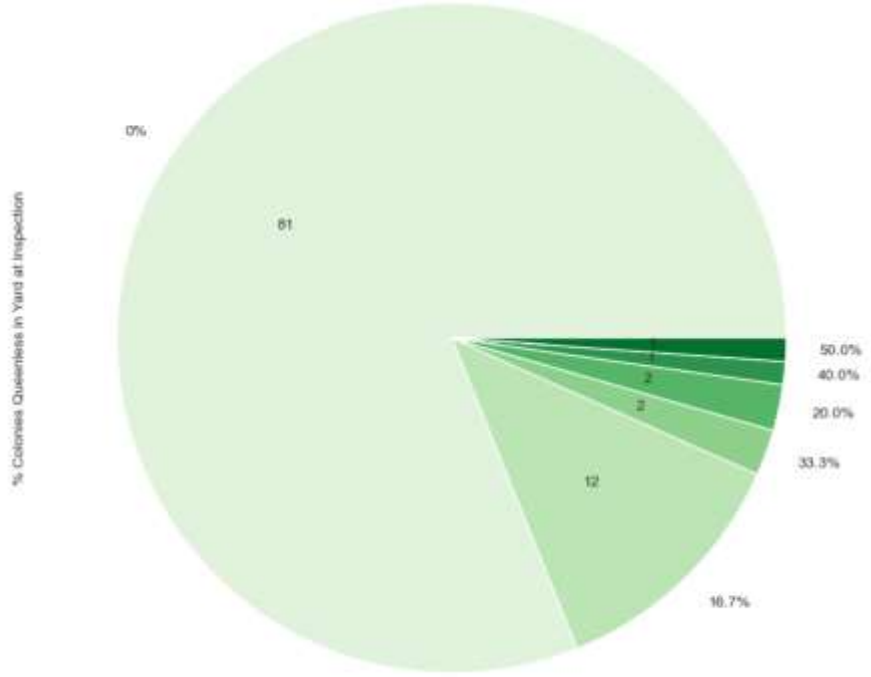
Dataset correlation - 2019

```
# 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()
```

Dataset visualization - 2019





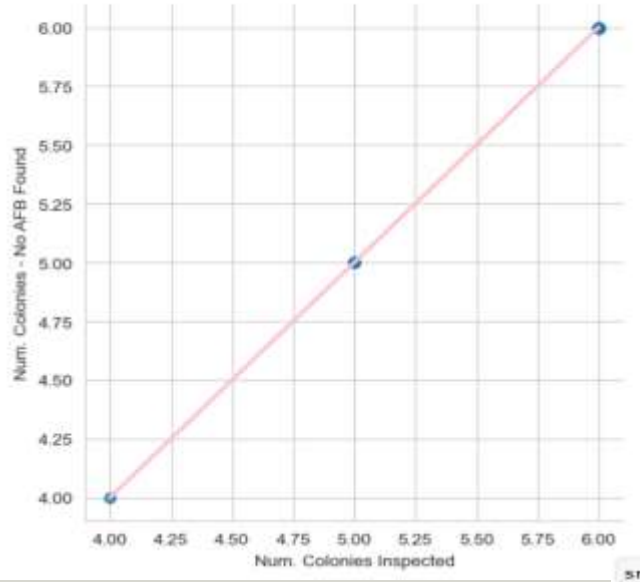



```
In [53]: df['County'].value_counts()

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

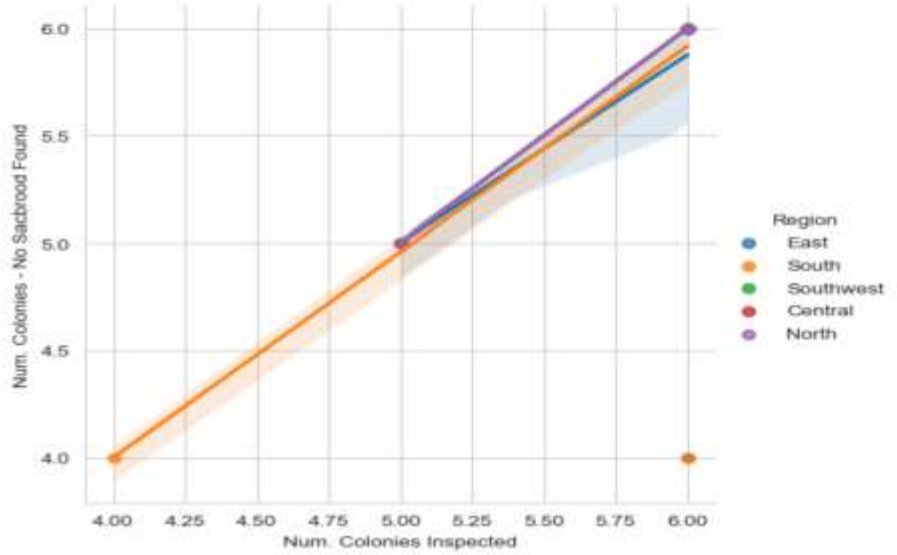
```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected',y ='Num. Colonies - No AFB Found', data = df, line_kws={'color': 'pink'})
```

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



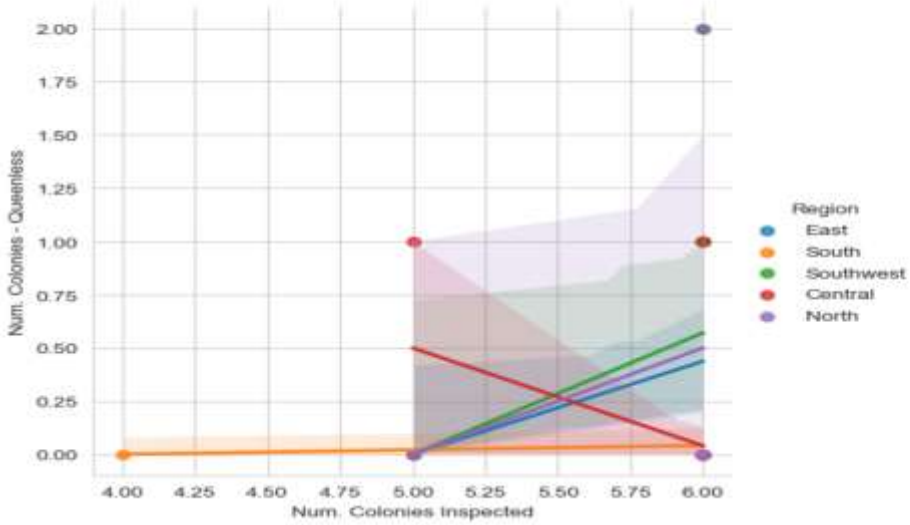

```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected', y ='Num. Colonies - No Sacbrood Found', data = df, hue ='Region')
```

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



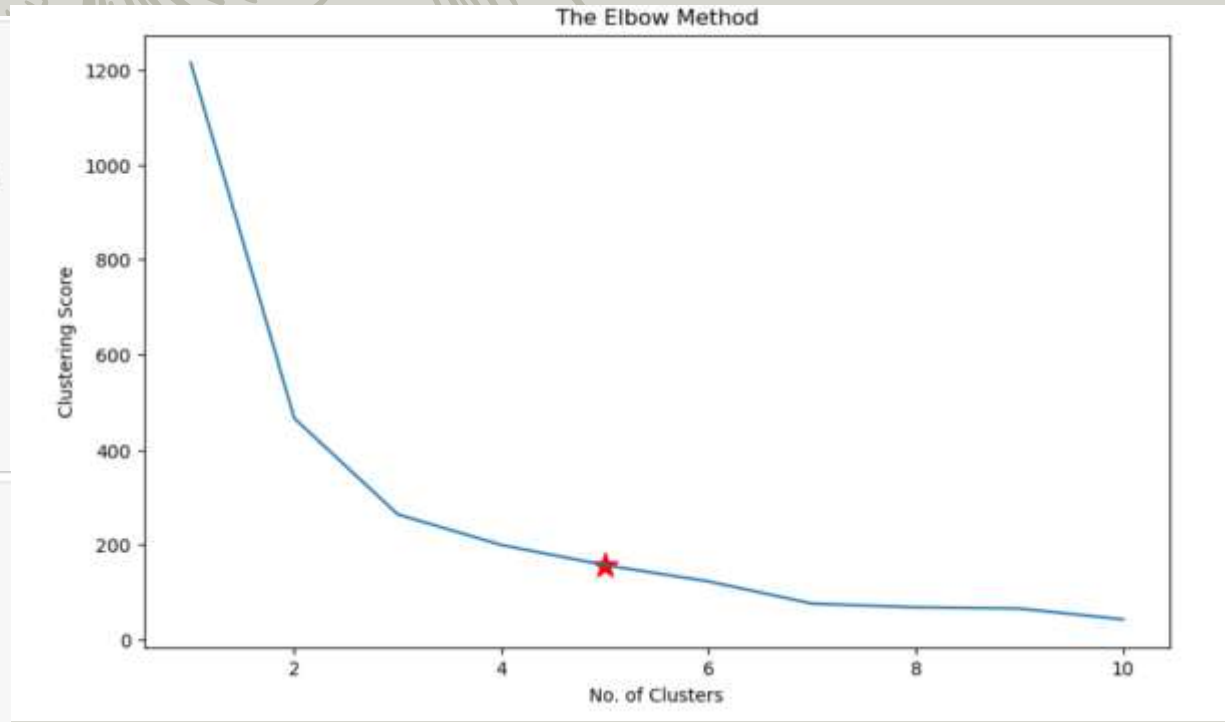

```
sns.set_style('whitegrid')
sns.lmplot(x ='Num. Colonies Inspected', y ='Num. Colonies - Queenless', data = df, hue ='Region')
```

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



Predictive analysis - K means clustering

```
clustering_score = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'random', random_state = 42)
    kmeans.fit(X)
    clustering_score.append(kmeans.inertia_) # inertia_ = Sum of squared distances of samples to their closest cluster center.


plt.figure(figsize=(10,6))
plt.plot(range(1, 11), clustering_score)
plt.scatter(5,clustering_score[4], s = 200, c = 'red', marker='*')
plt.title('The Elbow Method')
plt.xlabel('No. of Clusters')
plt.ylabel('Clustering Score')
plt.show()
```



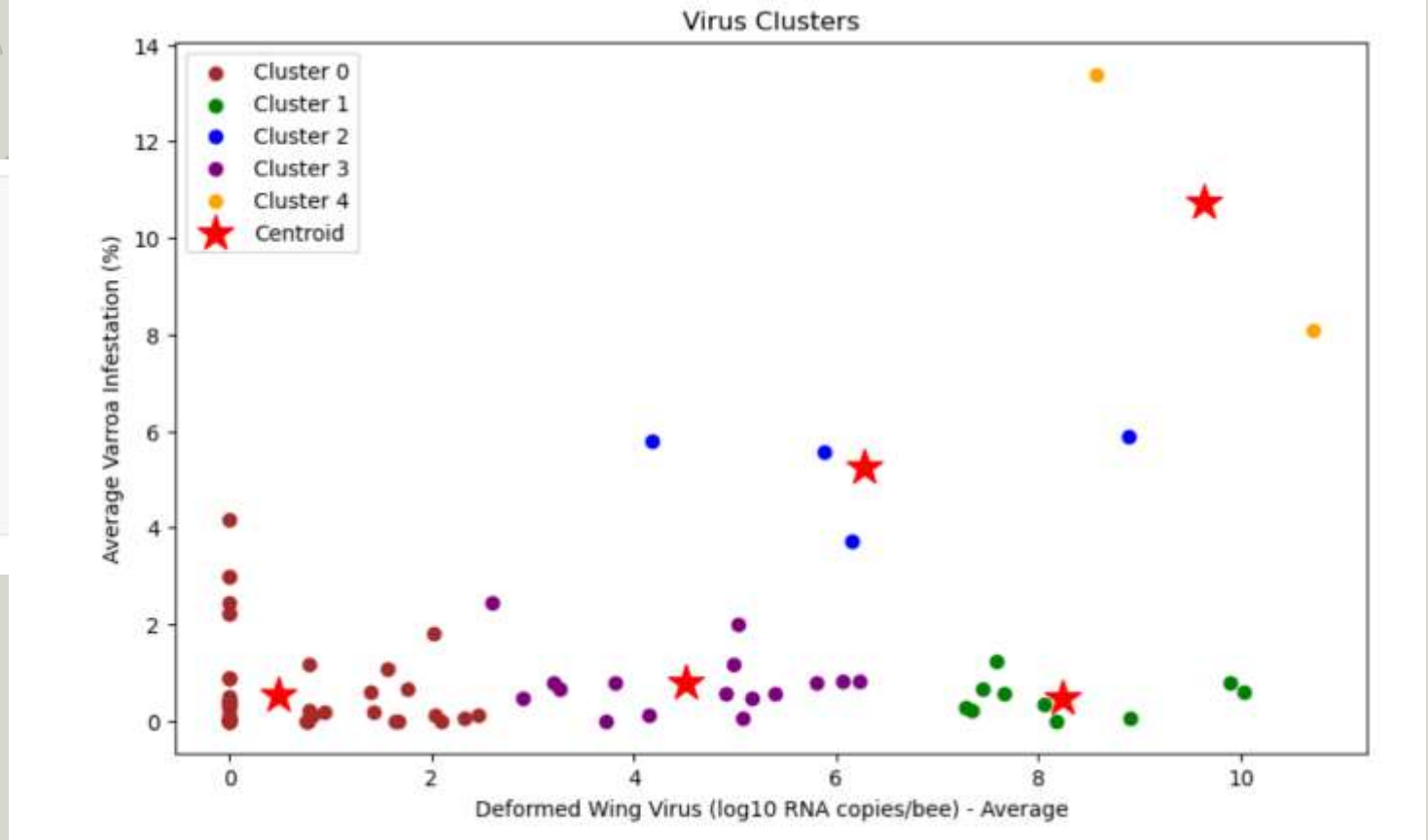

```
kmeans= KMeans(n_clusters = 5, random_state = 42)

# Compute k-means clustering
kmeans.fit(X)

# Compute cluster centers and predict cluster index for each sample.
pred = kmeans.predict(X)

pred
```

```
array([0, 0, 3, 0, 0, 3, 1, 1, 3, 3, 1, 0, 0, 2, 0, 3, 1, 0, 0, 0, 0, 0,
       0, 0, 2, 0, 0, 0, 0, 0, 1, 0, 3, 2, 0, 0, 3, 0, 3, 1, 0, 0, 0, 4,
       4, 0, 0, 3, 0, 0, 0, 1, 1, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 3, 0, 0, 3, 3, 0, 0, 0, 0, 3, 1, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0,
       0, 3])
```

Predictive analysis - K means clustering

```
plt.figure(figsize=(10,6))
plt.scatter(X[pred == 0, 0], X[pred == 0, 1], c = 'brown', label = 'Cluster 0')
plt.scatter(X[pred == 1, 0], X[pred == 1, 1], c = 'green', label = 'Cluster 1')
plt.scatter(X[pred == 2, 0], X[pred == 2, 1], c = 'blue', label = 'Cluster 2')
plt.scatter(X[pred == 3, 0], X[pred == 3, 1], c = 'purple', label = 'Cluster 3')
plt.scatter(X[pred == 4, 0], X[pred == 4, 1], c = 'orange', label = 'Cluster 4')

plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:, 1],s = 300, c = 'red', label = 'Centroid', marker='*')

plt.xlabel('Deformed Wing Virus (log10 RNA copies/bee) - Average')
plt.ylabel('Average Varroa Infestation (%)')
plt.legend()
plt.title('Virus Clusters')
```

```
Text(0.5, 1.0, 'Virus Clusters')
```



Predictive analysis – Random forest classifier

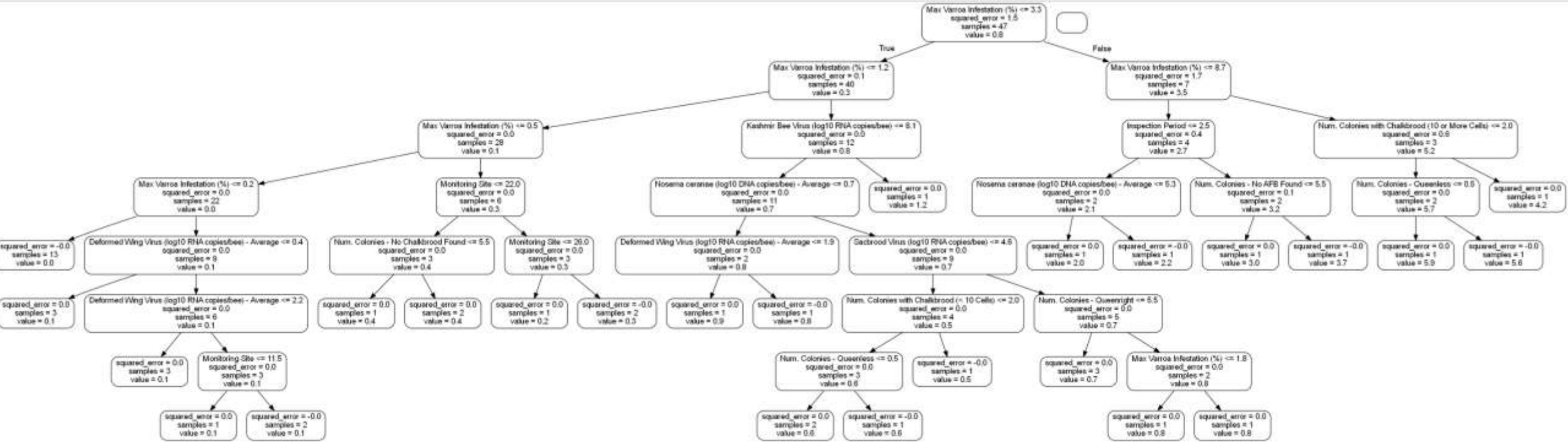
```
# Import the model we are using
from sklearn.ensemble import RandomForestRegressor

# Instantiate model
rf = RandomForestRegressor(n_estimators= 1000, random_state=42)

# Train the model on training data
clf = rf.fit(train_features, train_labels);
```

```
# Use the forest's predict method on the test data
predictions = rf.predict(test_features)

# Calculate the absolute errors
errors = abs(predictions - test_labels)

# Print out the mean absolute error (mae)
print('Mean Absolute Error:', round(np.mean(errors), 2), 'degrees.')
```

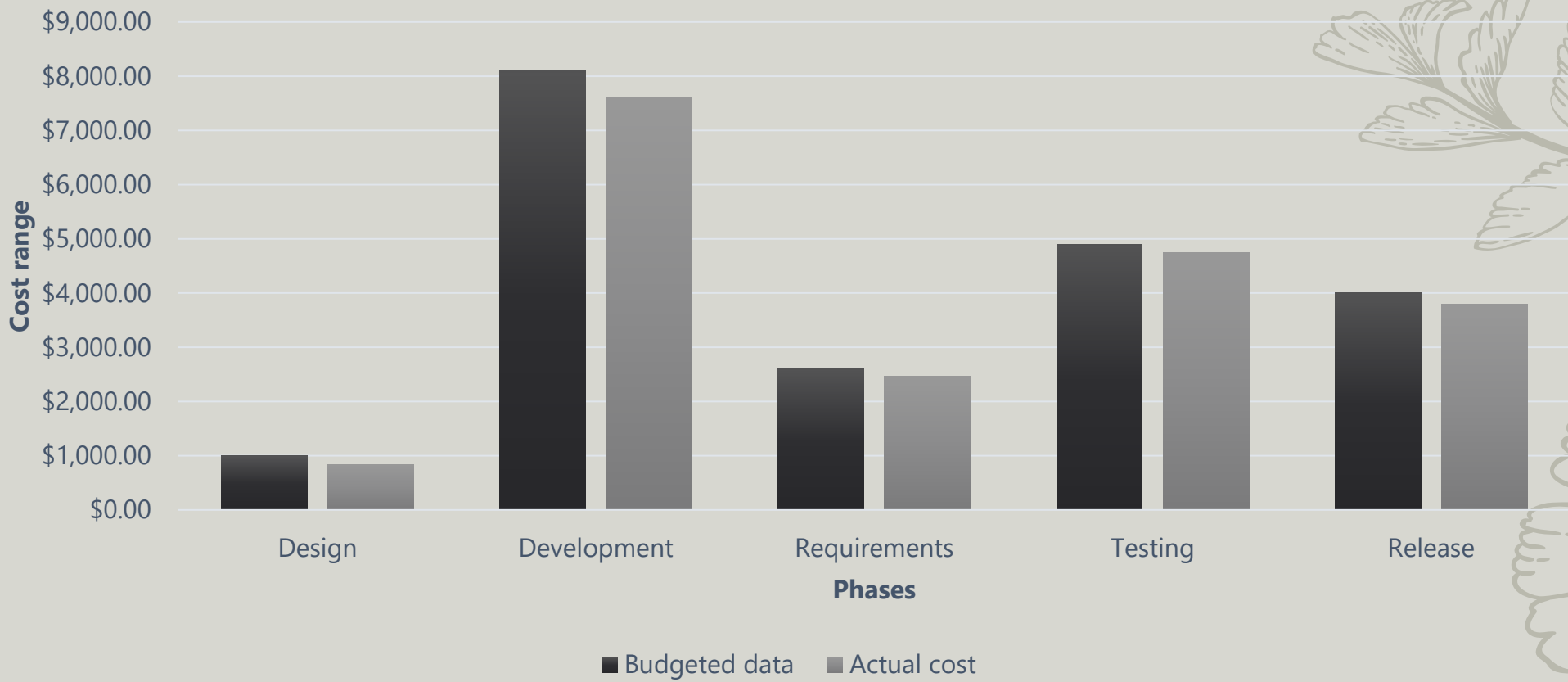
```
Mean Absolute Error: 0.22 degrees.
```

```
# Calculate and display accuracy
clf.score(train_features, train_labels)
```

```
0.9636324831180688
```



Project Cost



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

Update(25'jan): 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.

Update(08'Mar): Testing the clustering model is challenging

Update(15'Mar): No challenges

Update(29'Mar): Faced challenges executing Random Forest

Update (10'April): No challenges

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)

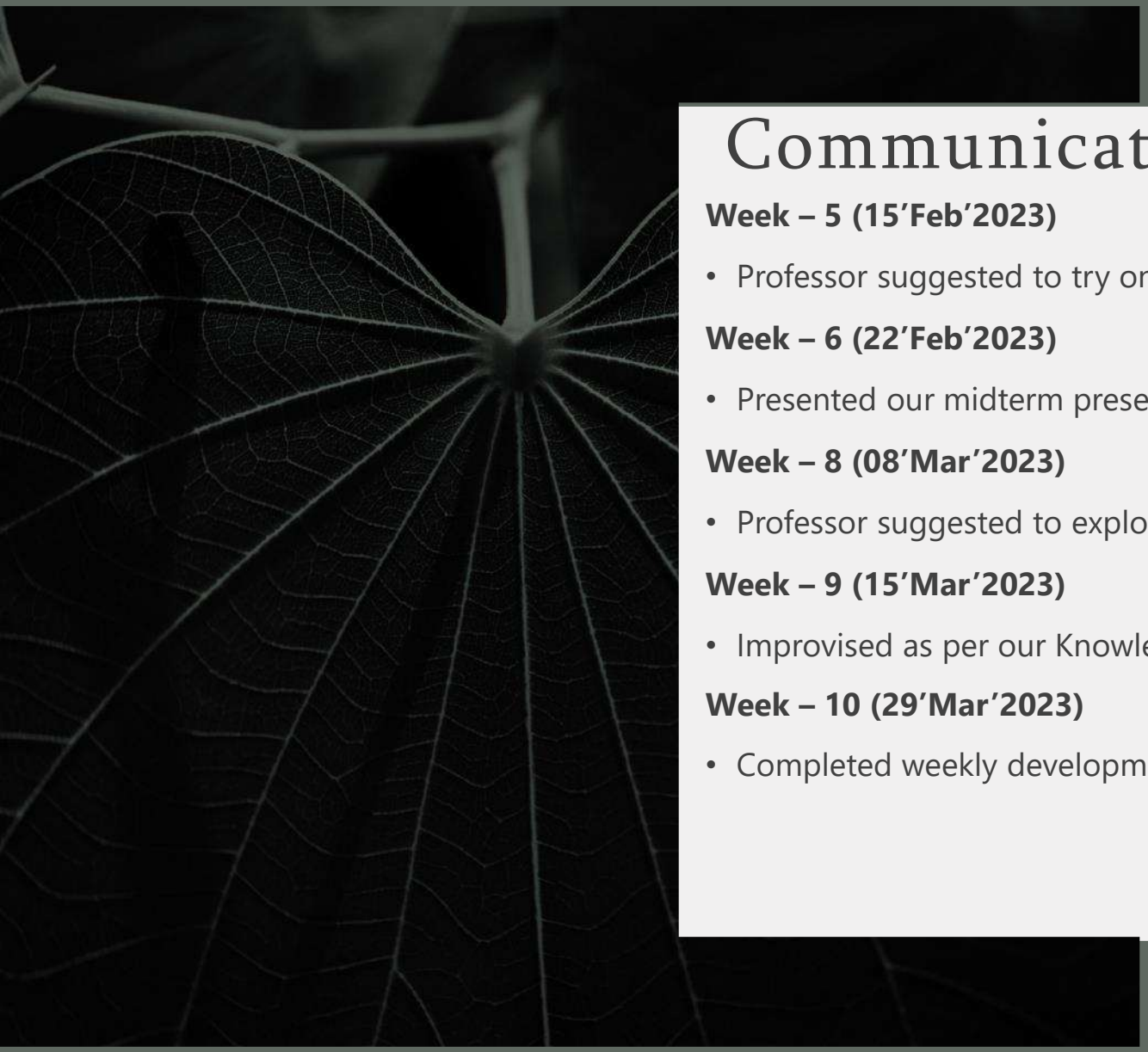
- Professor provided tutorial sites for building a predictive model
- Professor suggested to learn these models and understand clustering algorithms

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



Communications (Continued)

Week – 5 (15’Feb’2023)

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

Week – 6 (22’Feb’2023)

- Presented our midterm presentation

Week – 8 (08’Mar’2023)

- Professor suggested to explore on Random Forest predictive analysis

Week – 9 (15’Mar’2023)

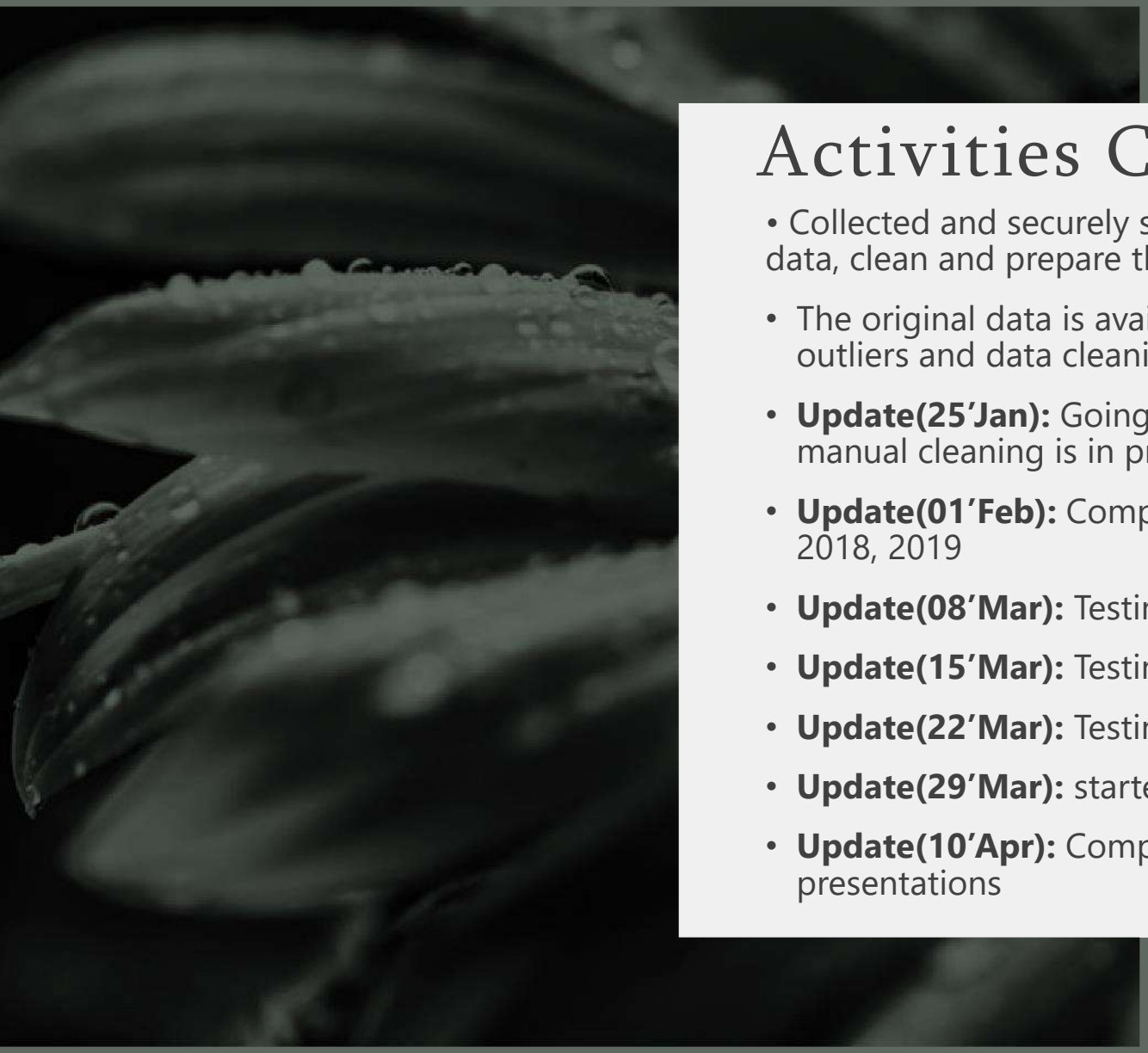
- Improvised as per our Knowledge

Week – 10 (29’Mar’2023)

- Completed weekly developmental progress and got feedbacks from Professor

Team meetings

Date	Agenda	Budgeted hours	Attendees	Approval of previous minutes
29/03/2023	Weekly status update – week 10	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
15/03/2023	Weekly status update – week 9	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
08/03/2023	Weekly status update – week 8	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
22/02/2023	Weekly status update – week 6	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
15/02/2023	Weekly status update – week 5	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
08/02/2023	Weekly status update – week 4	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
01/02/2023	Weekly status update – week 3	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
25/01/2023	Weekly status update – week 2	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved
18/01/2023	Weekly status update – week 1	0.15	1. Moganaviniith Rathinavel 2. Paras Kishorbhai Gangani 3. Ragavi Mudaliyar	Approved



Activities Completed This week

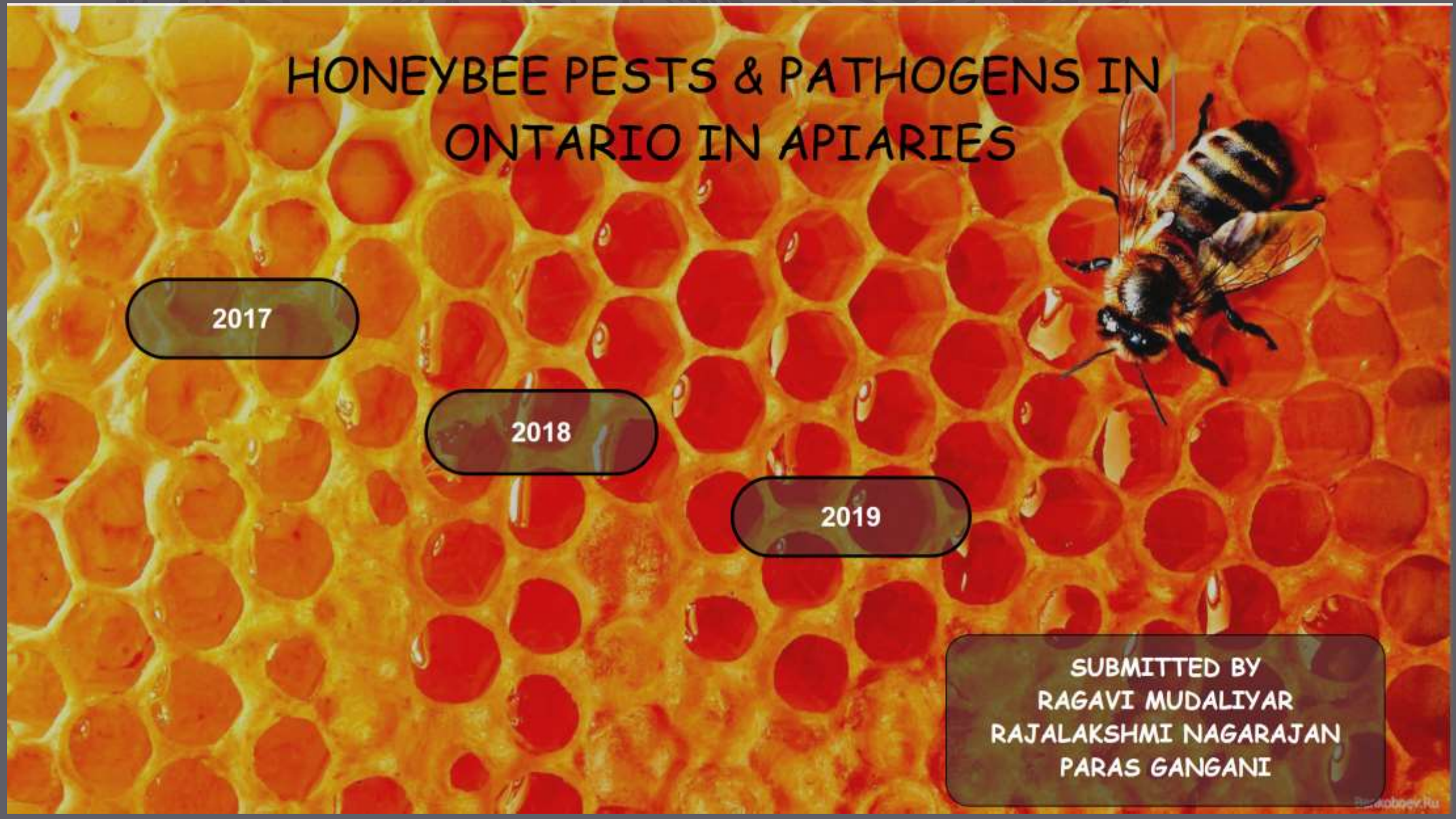
• Collected and securely stored the original data. Using copies of the original data, clean and prepare the data for analysis

- 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'Mar):** Testing the clustering model
- **Update(15'Mar):** Testing on Random Forest
- **Update(22'Mar):** Testing Completed on Random Forest
- **Update(29'Mar):** started working on visualizations
- **Update(10'Apr):** Completed the visualizations and ready for final presentations

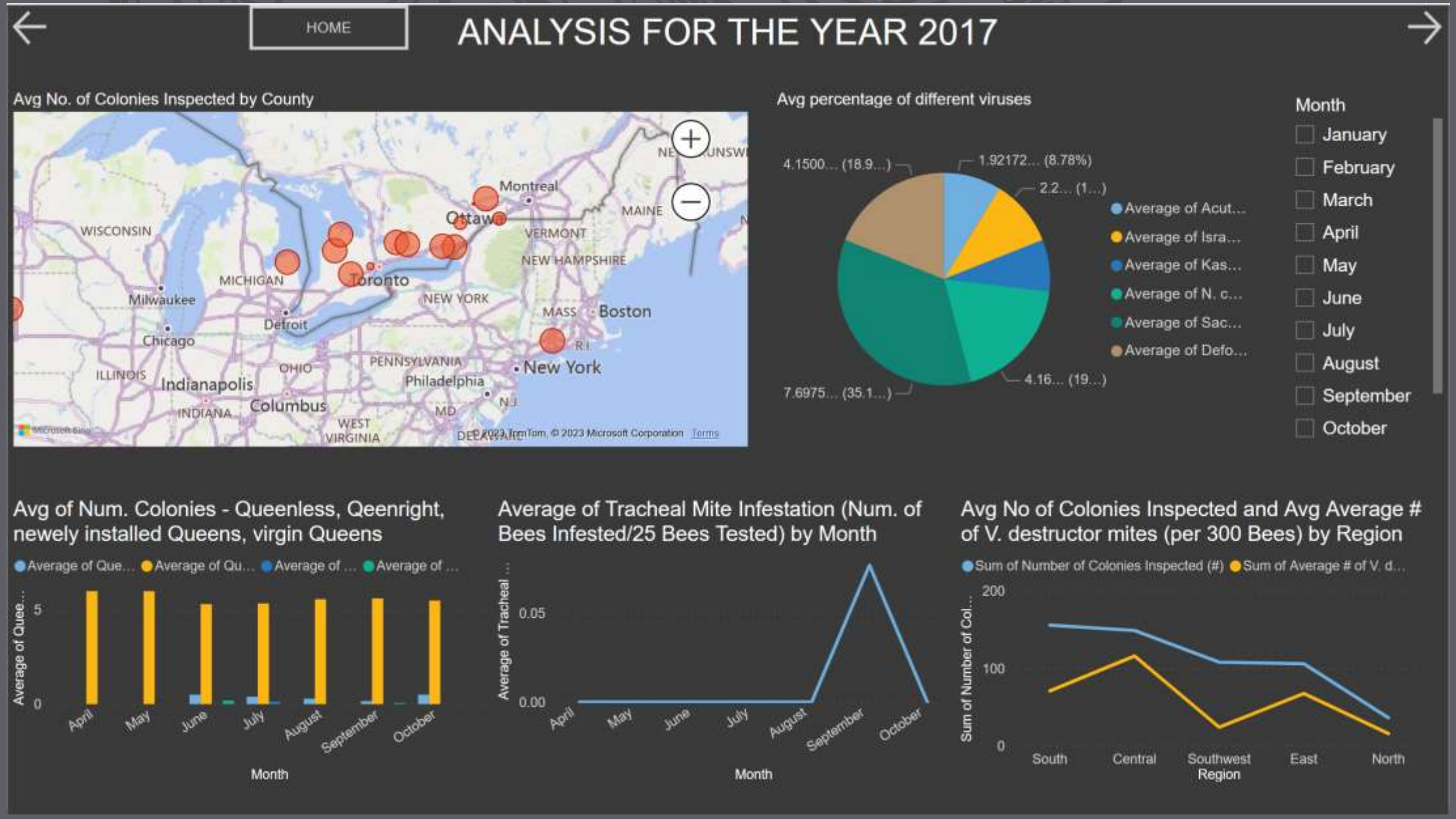
Activities to be Completed Before Next Report

- 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'Mar):** Work on testing
- **Update(15'Mar):** Complete the testing
- **Update(22'Mar):** Start working on visualizations
- **Update(29'Mar):** Complete designing dashboard
- **Update(10'Apr): Completed the project**

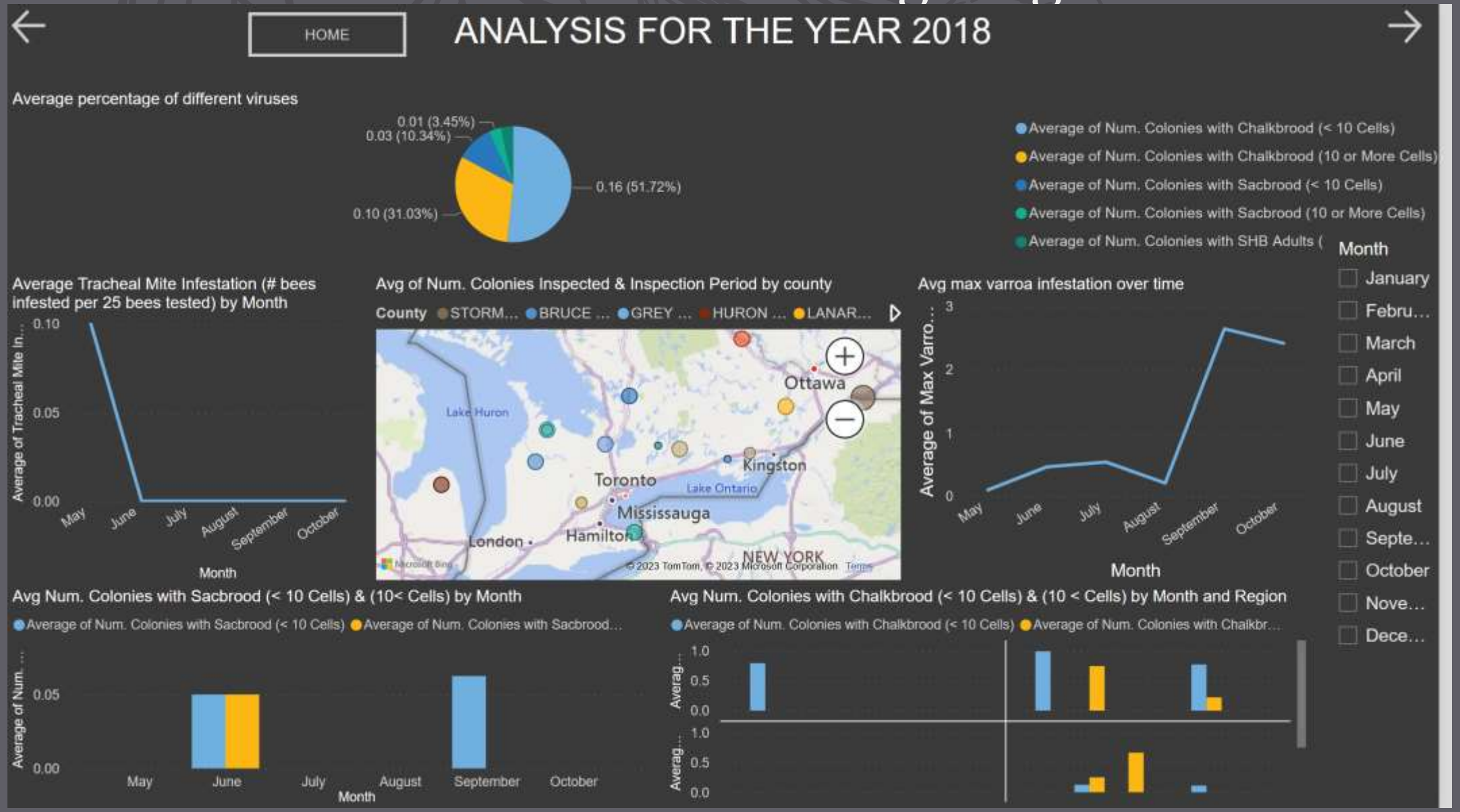
PowerBI Dashboard Designing – Landing Page



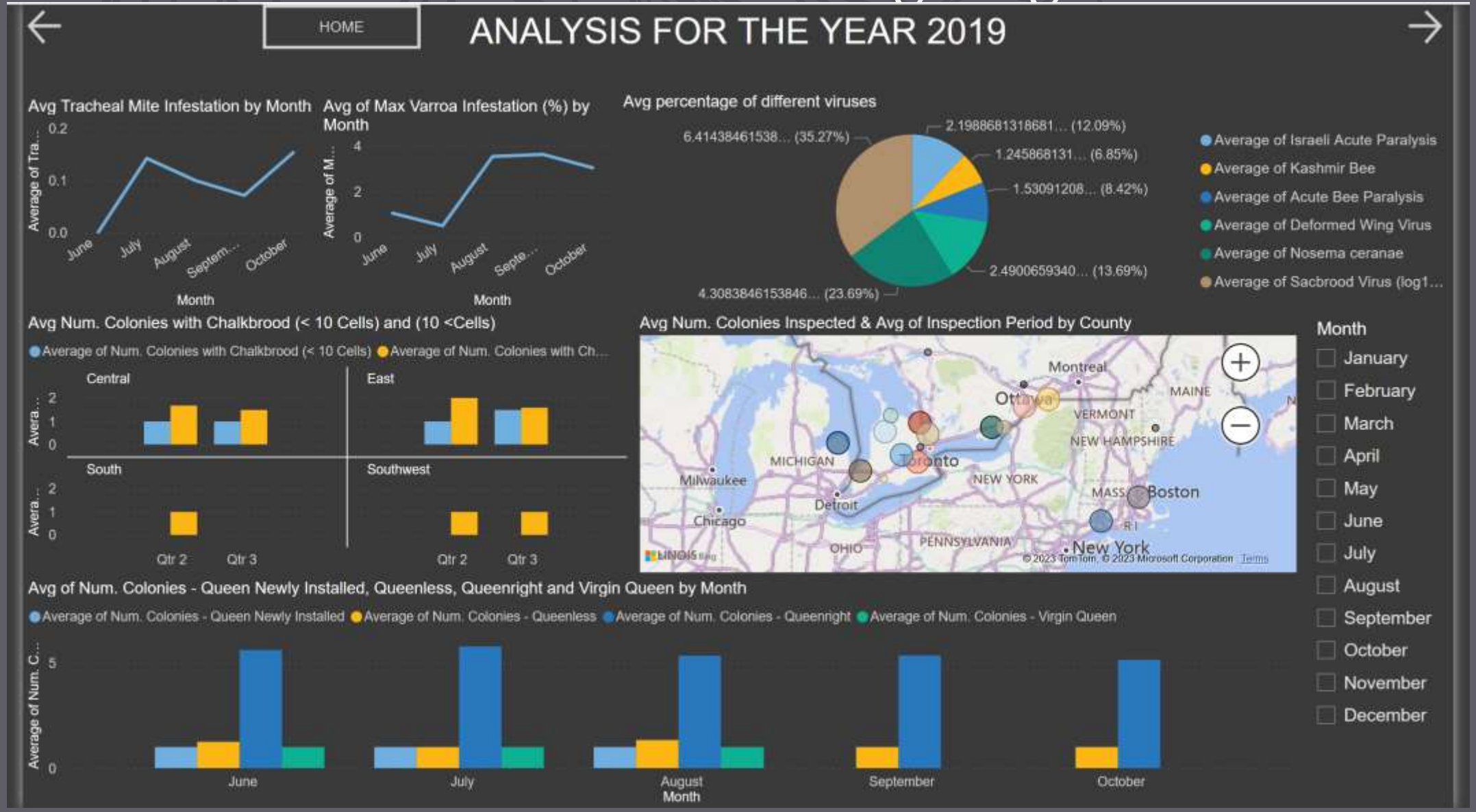
PowerBI Dashboard Designing – 2017



PowerBI Dashboard Designing – 2018



PowerBI Dashboard Designing – 2019



PowerBI Dashboard Publish link:

https://app.powerbi.com/groups/me/reports/36951c4c-cbc1-4bda-bb8a-7bae7135367d/ReportSection02c1b0c8175b44ac4109?ctid=da9a94b6-4681-49bc-bd7c-bab9eac0ad3c

Thank you