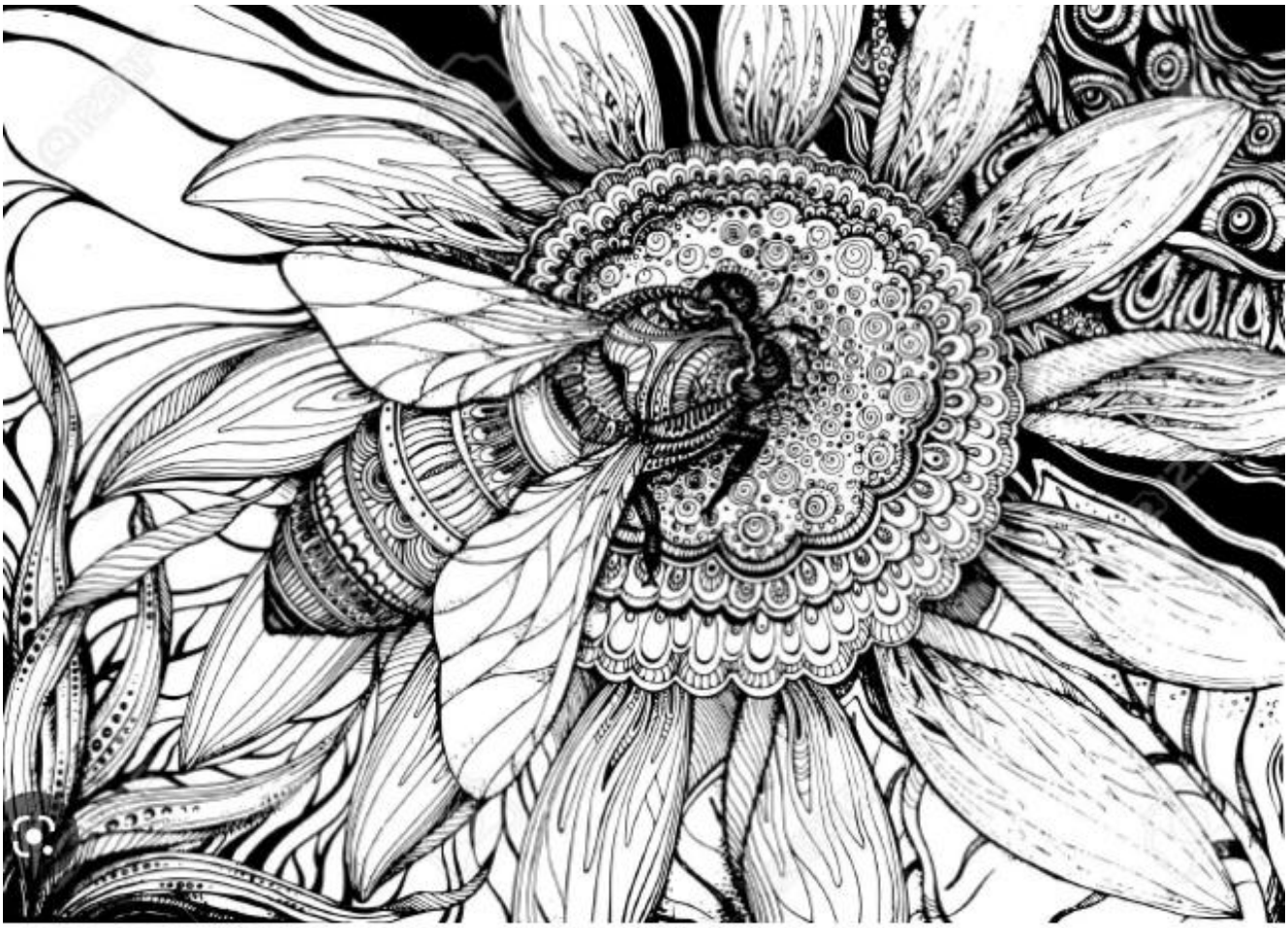
Survival Analysis Report: Bee Mortality Dataset

Author: Syed Bokhari

Date of Submission: 7th February, 2023



Introduction:

Bees are an important part of the global food production. Bees are pollinators which play a key role in sustaining food production systems. However, over the past few years, bee populations are suffering from declining numbers which does not bode well. It is believed that pesticides are the primary cause of this decline in bee populations. Among pesticides, herbicides are a category of chemicals that are considered safe for bees by regulatory authorities. However, the datasets used in this analysis indicate that is far from the case.

The datasets were acquired from https://qubeshub.org/publications/2989/2. The primary inspiration for the project as well as the source of the data used in the project is the paper Straw EA, Carpentier EN, Brown MJF. Roundup causes high levels of mortality following contact exposure in bumble bees. J Appl Ecol. 2021;58:1167– 1176.

The datasets used in the project were a result of experiments run by the researchers of the aforementioned paper. They sprayed various herbicides on bumblebees and measured the impact of the herbicides on the survival of the bees.

The datasets have the following attributes:

Microcolony (string): Refers to the box in which bees were sprayed. Although lettering is repeated across experiments, it is unrelated.

Treatment (string): Treatment used on the bees. Includes control group which received no treatment and different herbicides.

Origin Colony Name (string): Colony name from which the batch of bees was taken.

Time (int): Time elapsed in minutes till experiment start for the batch of bees.

Event(float): A censuring measure, alternating between 1 and 0. 1 indicates the death event occurred. 0 indicates it did not occur.

Objectives:

This project aims to use the data provided in the datasets for a Survival Analysis. The Survival Analysis will record the effects of various herbicides on the population of Bumblebees over time. These results will be used to determine which if any of the herbicides used in the experiment are actually harmful to bumblebees or not.

Data Exploration:

Three datasets corresponding to three experiments with different herbicides were used for the project. Although there is data corresponding to a total of 5 experiments in the original paper, only3 experiments were chosen as they were relevant to the project. The other two experiments used chemically weaker formulations of the herbicides which was not considered relevant for this project.

Initially some Exploratory Data Analysis (EDA) was performed. The datasets were loaded into Pandas dataframes, which make for convenient manipulation of data for further analysis.

```
df1 = pd.read_excel('Data Exp1.xlsx')
```

```
df4 = pd.read_excel('Data Exp4.xlsx')
```

```
df5 = pd.read_excel('Data Exp5.xlsx')
```

Afterwards, the shape of the datasets was examined.

```
df1.shape
```

```
(161, 5)
```

```
df4.shape
```

```
(107, 5)
```

```
df5.shape
```

```
(99, 5)
```

The first dataset had 161 rows. The second dataset had 107 rows. The third dataset had 99 rows. All datasets had 5 columns. These were examined as well.

```
df1.head()
```

	MicroColony_ID	Treatment	OriginColony_name	Time	Event
0	a	Control	Colony 1	1440.0	1.0
1	b	Control	Colony 1	1440.0	0.0
2	c	Control	Colony 1	1440.0	0.0
3	d	Control	Colony 1	1440.0	0.0
4	e	Control	Colony 1	1440.0	0.0

```
df4.head()
```

	MicroColony_ID	Treatment	OriginColony_name	Time	Event
0	a	Control	Colony 3	1440.0	1.0
1	a	Control	Colony 3	1440.0	0.0
2	a	Control	Colony 3	1440.0	0.0
3	a	Control	Colony 3	1440.0	0.0
4	a	Control	Colony 3	1440.0	0.0

```
df5.head()
```

	MicroColony_ID	Treatment	OriginColony_name	Time	Event
0	a	Control	Colony 9	1440.0	0.0
1	a	Control	Colony 9	1440.0	0.0
2	a	Control	Colony 9	1440.0	0.0
3	a	Control	Colony 9	1440.0	0.0
4	a	Control	Colony 9	1440.0	0.0

It was noted that each of the datasets had the same name of columns indicating the same thing. This makes for convenient analysis.

Subsequently, the datasets were checked for missing values:

```
df1.isna().sum().sum()
```

```
0
```

```
df4.isna().sum().sum()
```

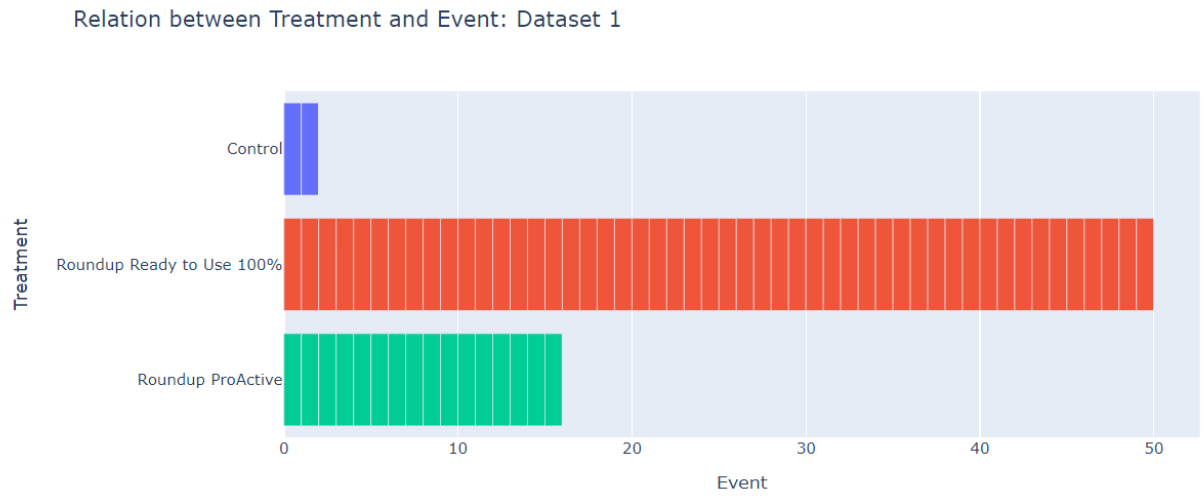
```
0
```

```
df5.isna().sum().sum()
```

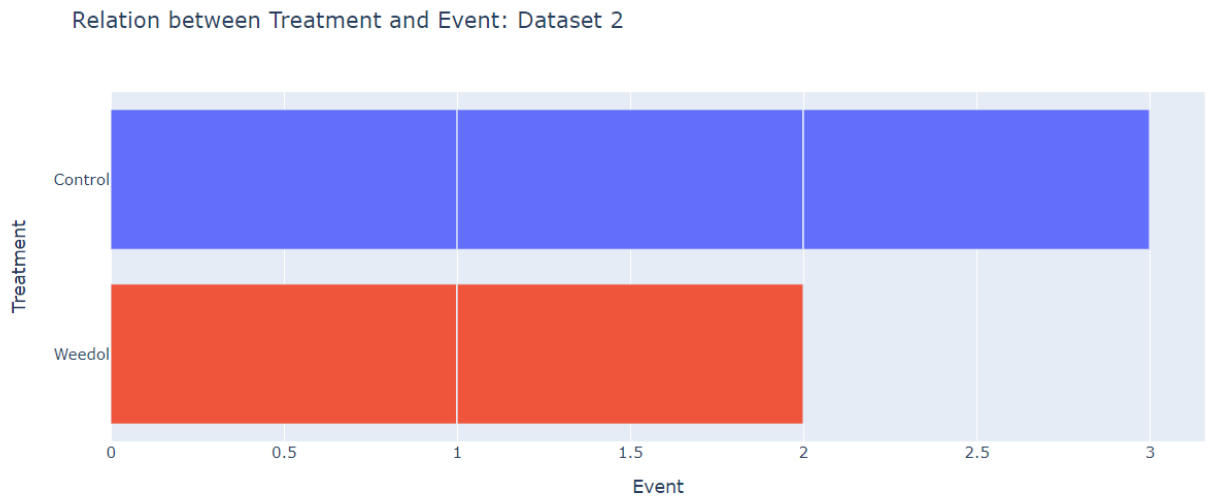
```
0
```

It turned out that there were no missing values in the datasets.

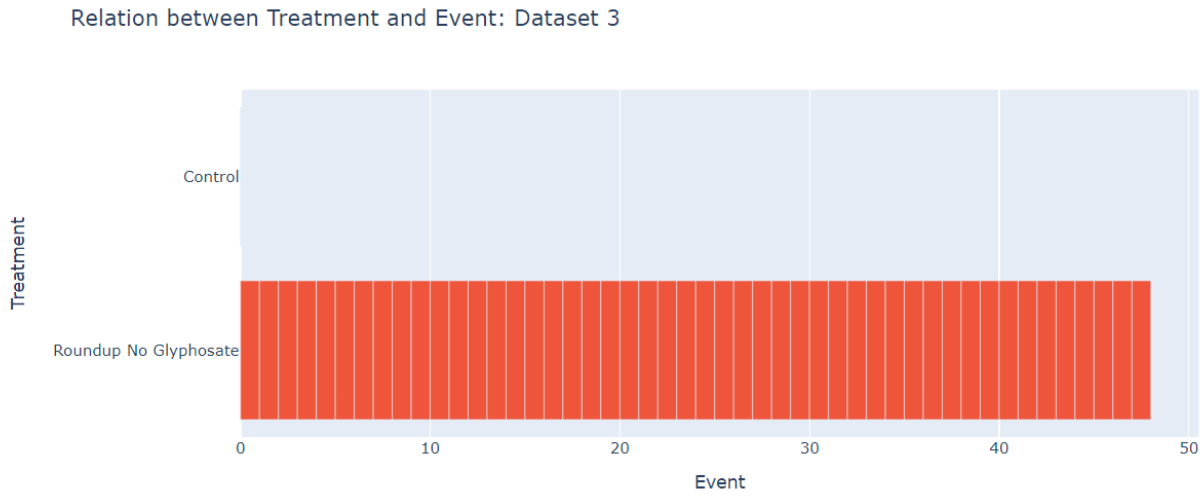
As each experiment examined the impact of treatment on event, the relation between treatment and event was plotted. For dataset1, this was as follows:



Clearly, there is some impact of herbicides on the event as noted by the bar graph. A similar plot was made for dataset 2:

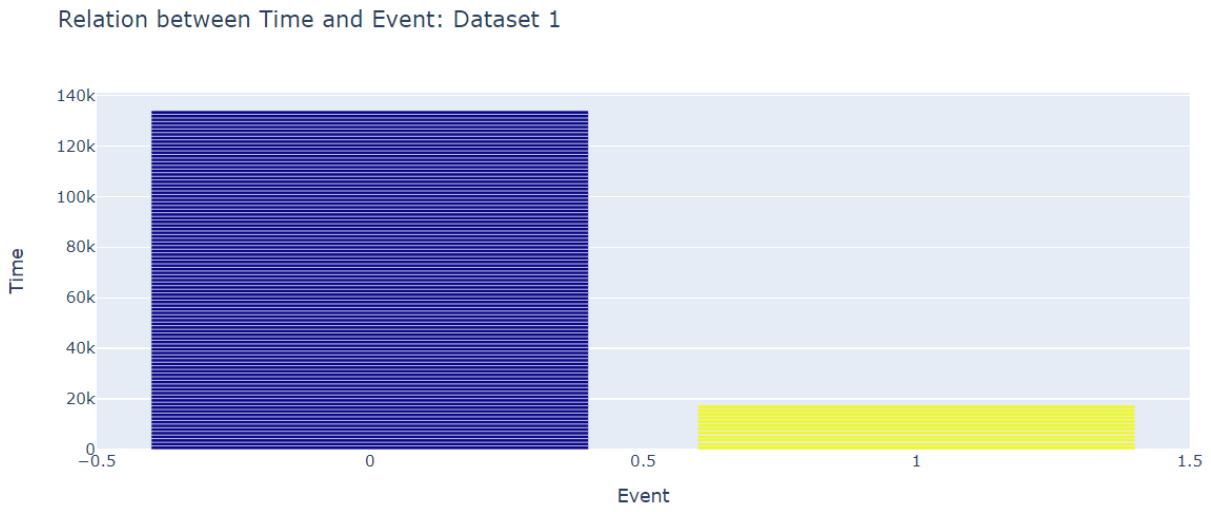


This plot is interesting because it indicates that while choice of herbicide did have an impact, it was in the opposite direction. This will bear further examination in later stages. Next, a plot for the final dataset on similar lines was created:

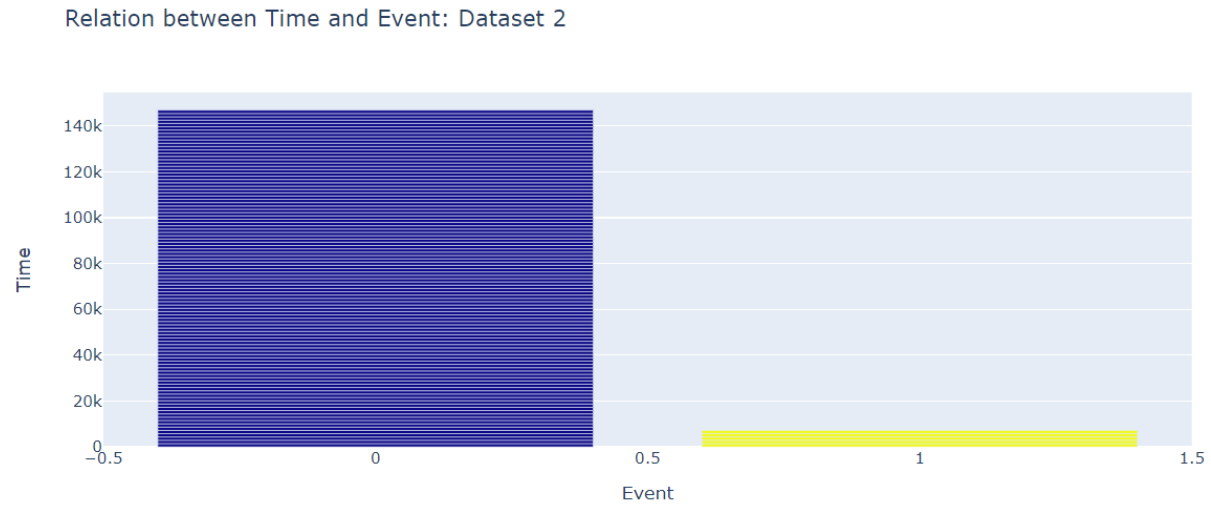


This plot is interesting too as it shows that almost all event occurrences for the experiment were due to herbicide. This will be looked at in detail in later stages o the project.

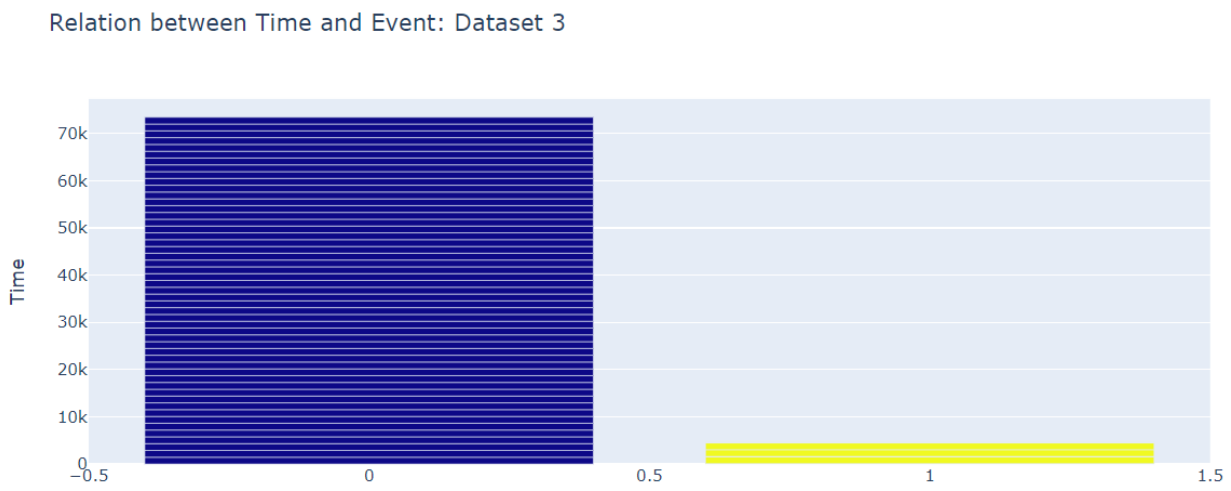
We also examine the relationship between Time and Event, two variables that will be modeled further onwards in the Survival Analysis. For dataset1, the plot is as follows:



Its clear that a relationship exists between time and event for this dataset. Moving on to dataset 2:



This time the relationship is less pronounced, although it still exists. Plotting this for dataset 3 yields the following chart:



There is little difference between this plot and the plot for dataset 2.

To summarize, EDA indicated that there was an apparent relationship between Event and Treatment or type of herbicide used. Event here refers to life or death of the bees. Similarly, a relationship also exists between Time and Event, where Time is time elapsed till start of experiment.

Plan:

With the datasets ready, the next step was to perform the Survival Analysis. This would be done by first plotting the Kaplan-Meier curve for the datasets. Then the Kaplan-Meier curve would be plotted for each Treatment type in each dataset to visually examine the impact of the treatment on survival probability.

Afterwards, relevant data from each dataset will be passed to a Cox Proportional Hazards model and results will be noted. Further discussion of results will follow.

Survival Analysis:

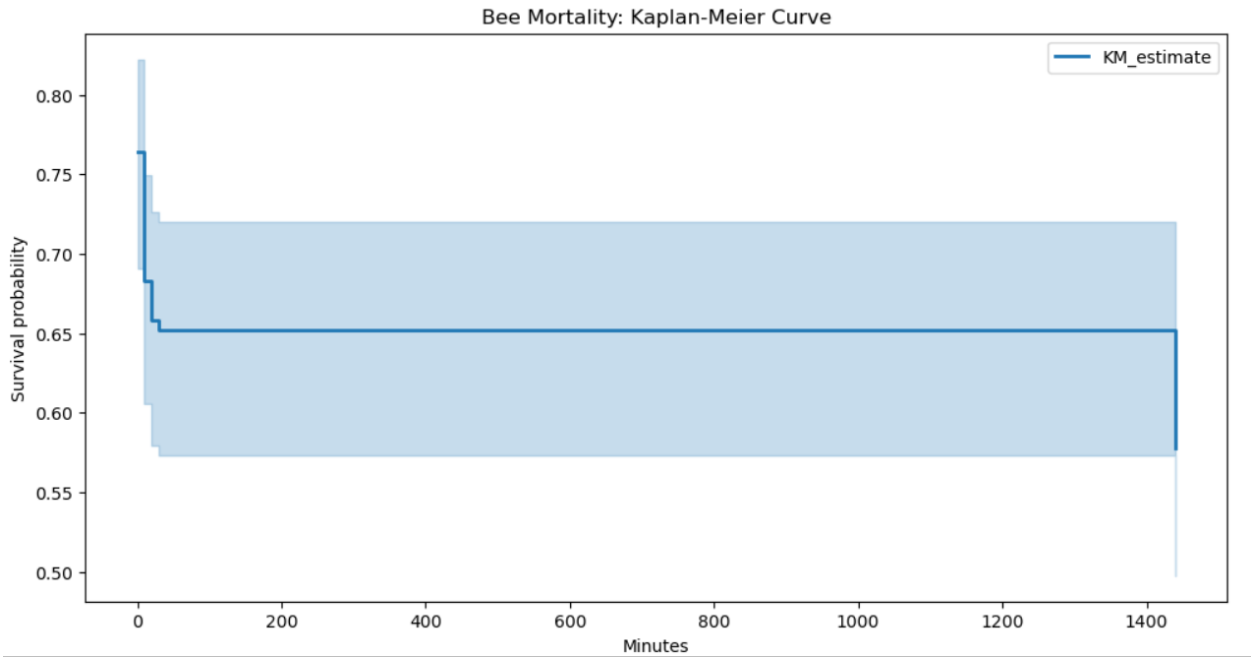
Dataset-1 Kaplan-Meier Curves:

Next step is to plot Kaplan-Meier curves to see the impact of treatment with herbicides on survival probability of bees. For dataset 1, this is done by fitting the Time and Event columns to a Kaplan-Meier Fitter object.

```
kmf = KaplanMeierFitter()
```

```
kmf.fit(df1.Time, df1.Event)
```

Next, the results are plotted for visual representation of survival probability:



The results clearly indicate that survival probability has some relation to time. The relatively flat line in the middle is because the dataset has very few unique values for time.

```
df1.Time.value_counts()
```

```
1440.0    105
0.0        38
10.0       13
20.0        4
30.0        1
Name: Time, dtype: int64
```

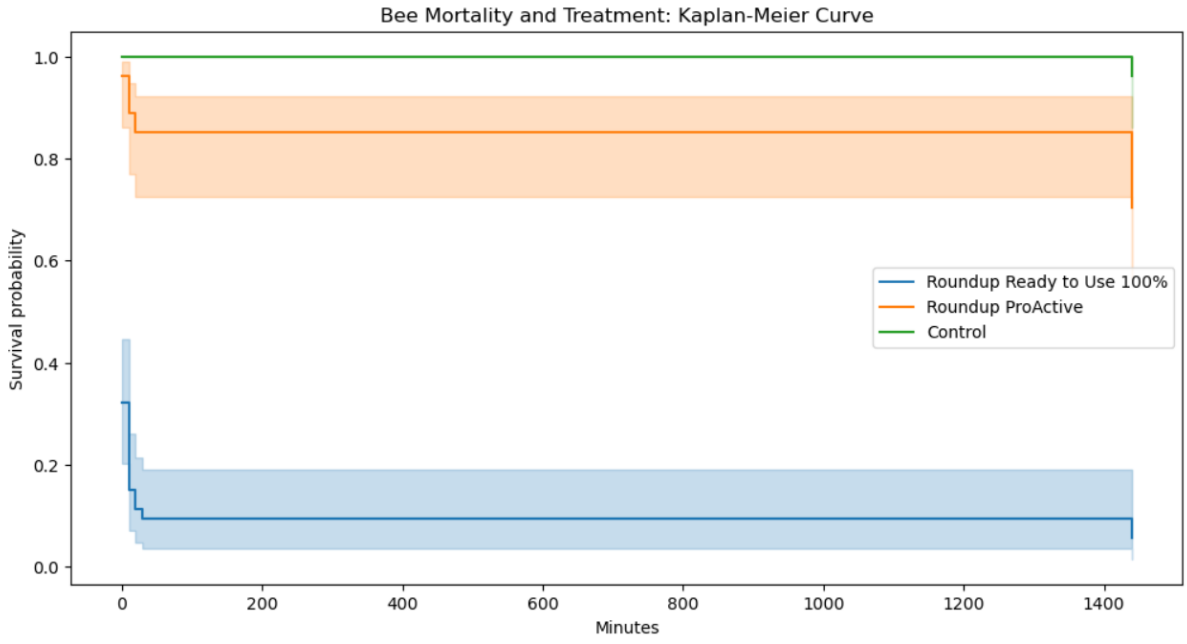
Next step is to conduct some feature engineering. The dataset is divided into smaller datasets containing only one of the treatment methods. The control treatment method refers to the control group for the experiment that was administered no herbicide. The other treatment groups were administered the specified herbicide.

```
df1['Treatment'].unique()
```

```
array(['Control', 'Roundup Ready to Use 100%', 'Roundup ProActive'],
      dtype=object)
```

```
df1a = df1[df1.Treatment=='Roundup Ready to Use 100%']
df1b = df1[df1.Treatment=='Roundup ProActive']
df1c = df1[df1.Treatment=='Control']
```

Kaplan-Meier curves are plotted for each of these datasets.



Its clear from this plot that treatment methods have an impact on survival probability for dataset1.

Dataset-1 Cox Proportional Hazards Model:

The data was next modeled with Cox Proportional Hazards model. This however required some feature engineering. The Cox Proportional Hazards model takes feature an event column, a time column and feature columns in one hot-encoded form. For datasets considered in the project only 1 feature column Treatment was present. This was one-hot encoded to prepare the dataset for modeling:

```
df_u = df1[['Treatment','Event']]
df_d = pd.get_dummies(df_u, drop_first=True)
df_d['Time'] = df1.Time
df_d.head()
```

	Event	Treatment_Roundup ProActive	Treatment_Roundup Ready to Use 100%	Time
0	1.0	0	0	1440.0
1	0.0	0	0	1440.0
2	0.0	0	0	1440.0
3	0.0	0	0	1440.0
4	0.0	0	0	1440.0

Next, the prepared data was modeled with Cox Proportional Hazards Model:

```
cph = CoxPHFitter()
cph.fit(df_d, duration_col='Time', event_col='Event')
cph.print_summary()
```

The results were as follows:

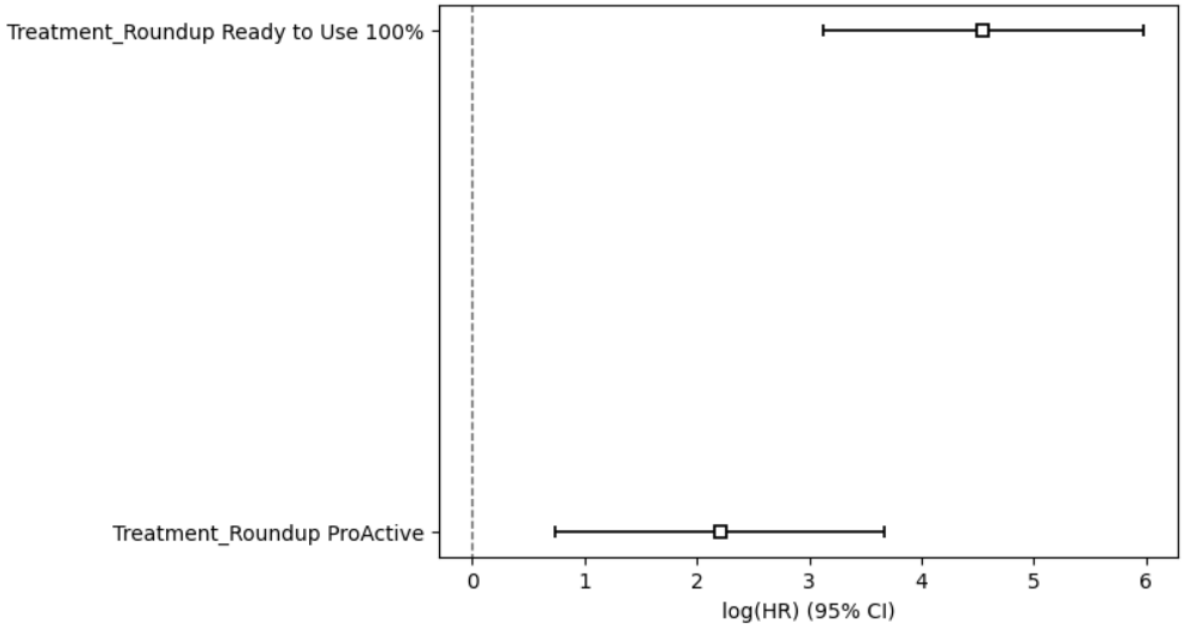
model	lifelines.CoxPHFitter
duration col	'Time'
event col	'Event'
baseline estimation	breslow
number of observations	161
number of events observed	68
partial log-likelihood	-262.50
time fit was run	2023-02-07 00:19:30 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Treatment_Roundup ProActive	2.20	9.06	0.75	0.73	3.67	2.08	39.42	0.00	2.94	<0.005	8.24
Treatment_Roundup Ready to Use 100%	4.55	94.60	0.73	3.12	5.98	22.62	395.59	0.00	6.23	<0.005	31.02

Concordance	0.88
Partial AIC	528.99
log-likelihood ratio test	132.71 on 2 df
-log2(p) of ll-ratio test	95.73

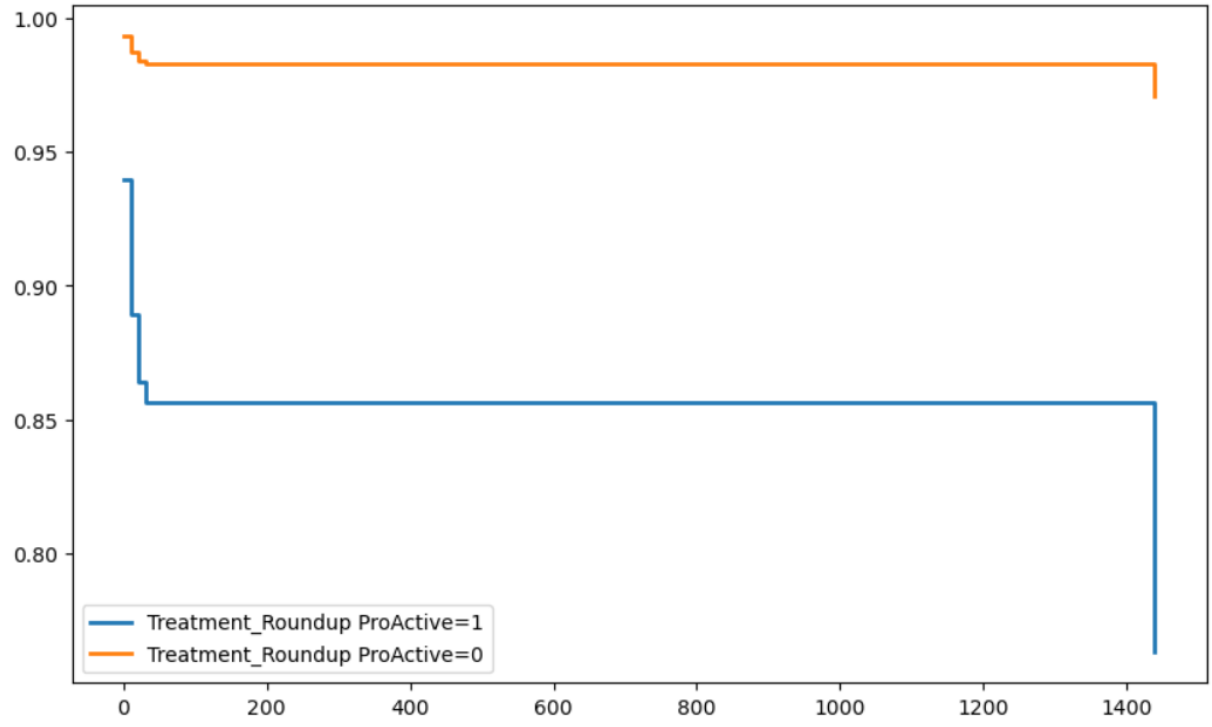
The results indicate that both Treatment types have an impact on survival as p-values for both are less than 0.005, indicating a significant result. Also, both have confidence

intervals that do not include 0 indicating a statistically significant relationship. To better visualize this, the confidence intervals were plotted:

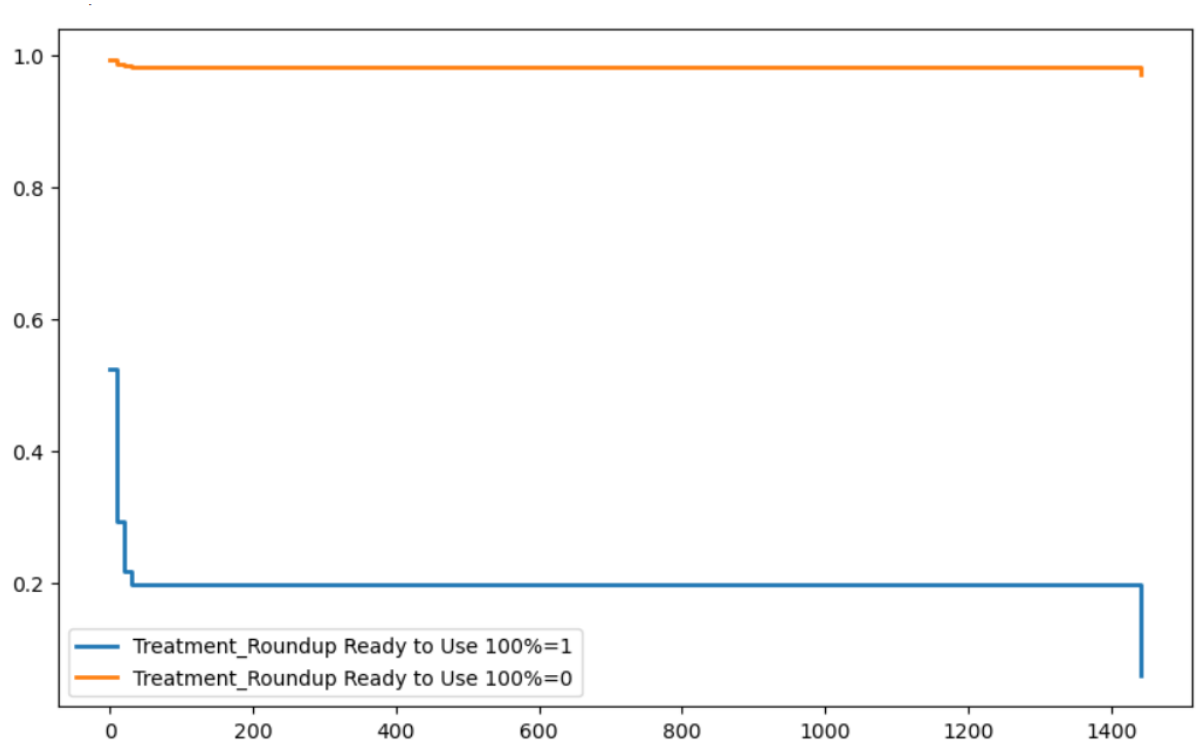


Clearly, 0 is not included in the confidence intervals so relationship is significant.

Next, the impact of the herbicides on the survival probability was explicitly plotted. For Roundup ProActive, the following plot was generated.



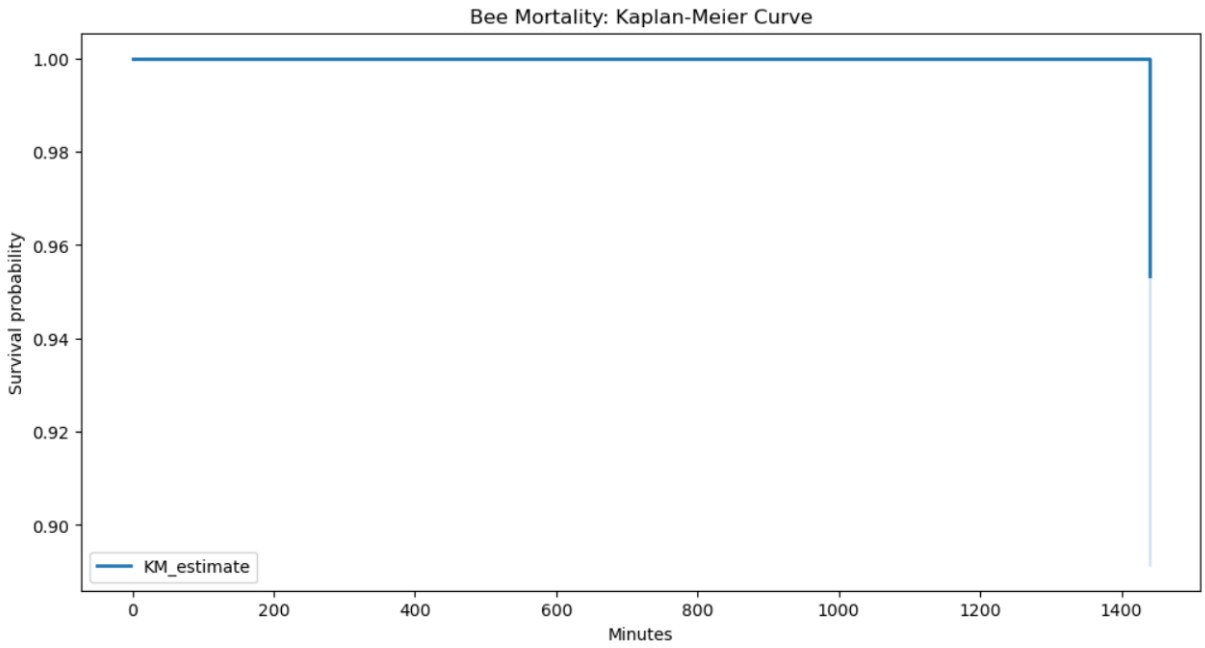
For Roundup Ready to Use 100%, the following plot was created:



The two plots indicate the explicit impact on survival probability of the two herbicides. It can be clearly observed that Roundup Ready to Use 100% has a far greater impact on survival probability than Roundup ProActive.

Dataset-2 Kaplan-Meier Curves:

Next, the Kaplan-Meier curves are plotted for dataset 2.

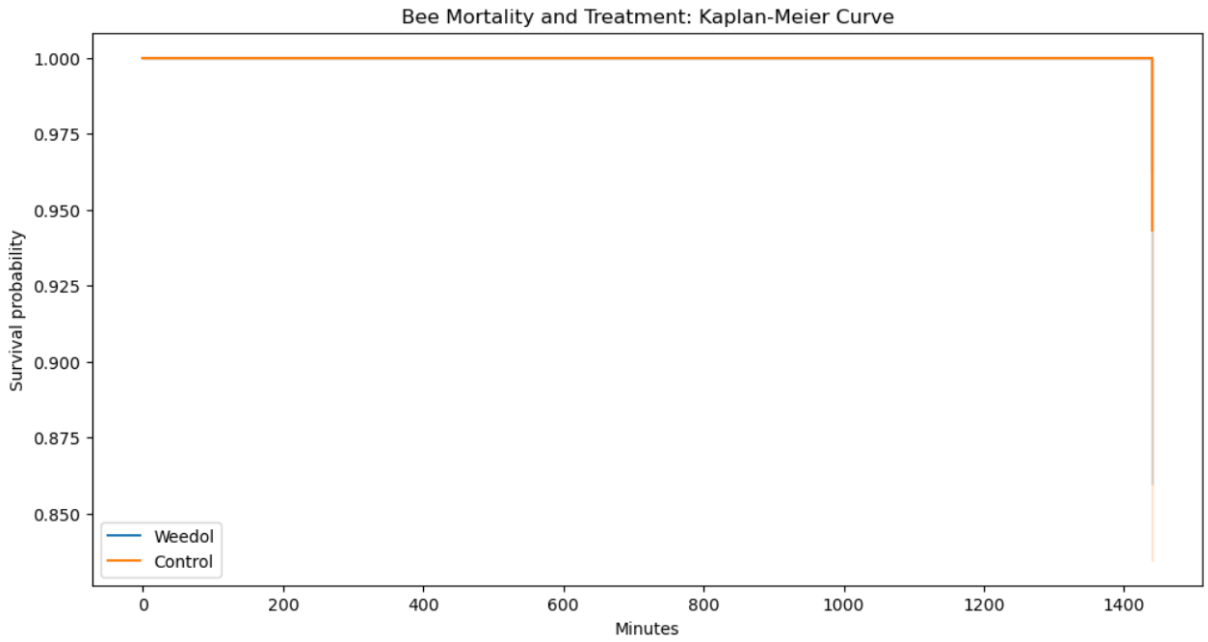


Note that the impact on survival probability is non-existent for the most part for this dataset. This is because there are very few death events ('1') in this dataset.

```
df4.Event.value_counts()
```

```
0.0    102
1.0      5
```

Once more, the data is divided according to Treatment type and then the resulting Kaplan-Meier curves are plotted.



The resulting curves are no different from the previous ones indicating there is little impact on survival probability due to the herbicide Weedol.

Dataset-2 Cox Proportional Hazards Model:

Next, the data is to be modeled using the Cox Proportional Hazards model, also referred to as Survival Regression. For this it is first modified so that the Treatment column which is a feature in the dataset is one-hot encoded.

```
df_u = df4[['Treatment','Event']]
df_d = pd.get_dummies(df_u, drop_first=True)
df_d['Time'] = df4.Time
df_d.head()
```

	Event	Treatment_Weedol	Time
0	1.0	0	1440.0
1	0.0	0	1440.0
2	0.0	0	1440.0
3	0.0	0	1440.0
4	0.0	0	1440.0

The data is then fitted for Survival Regression. The results are as follows:

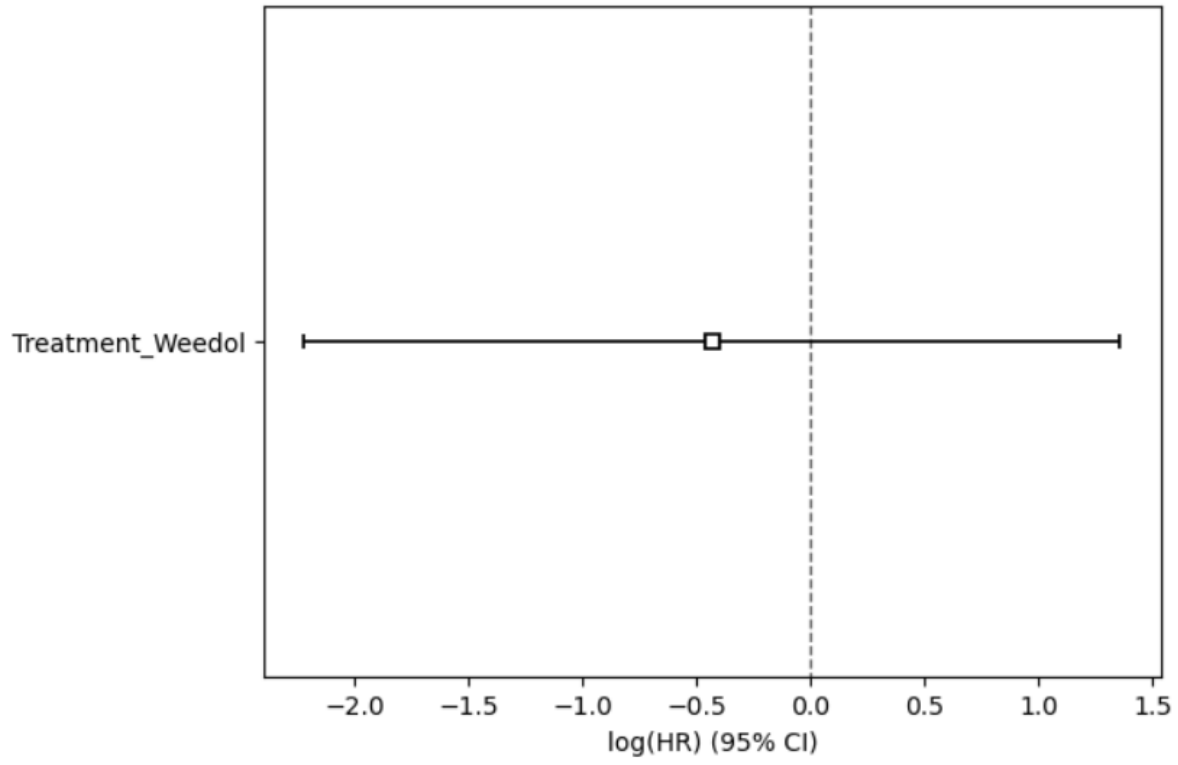
```
cph = CoxPHFitter()
cph.fit(df_d, duration_col='Time', event_col='Event')
cph.print_summary()
```

model	lifelines.CoxPHFitter
duration col	'Time'
event col	'Event'
baseline estimation	breslow
number of observations	107
number of events observed	5
partial log-likelihood	-23.15
time fit was run	2023-02-07 00:07:46 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Treatment_Weedol	-0.43	0.65	0.91	-2.22	1.36	0.11	3.88	0.00	-0.47	0.64	0.65

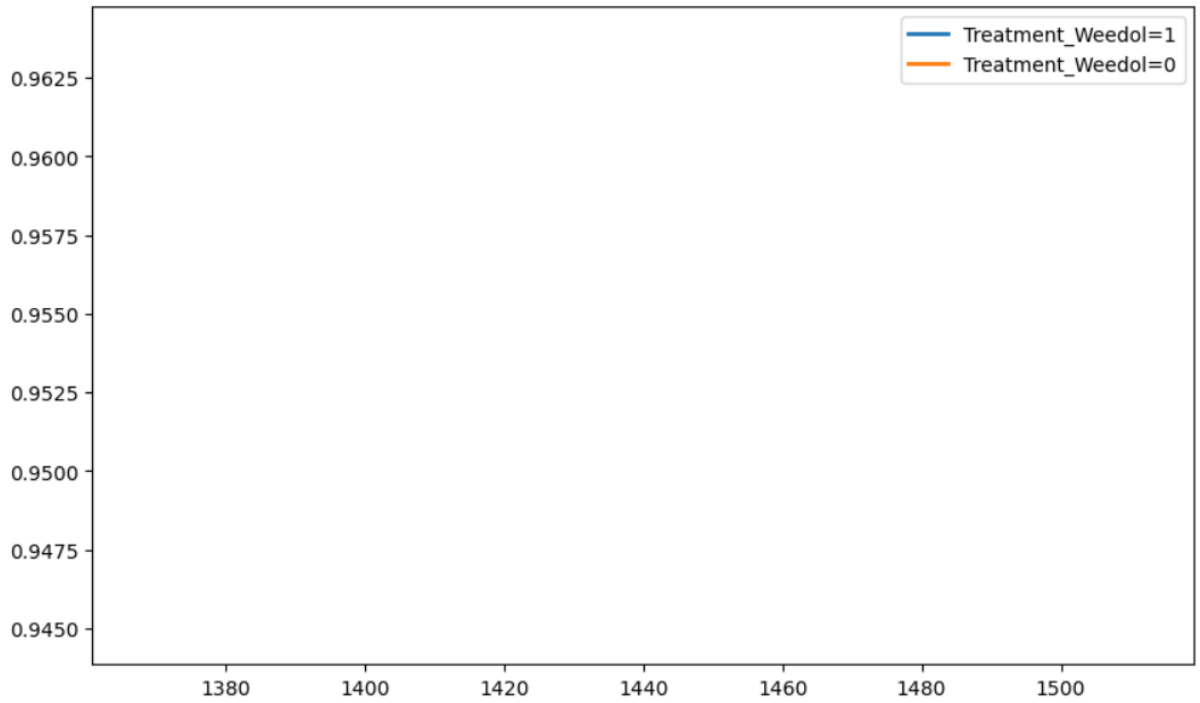
Concordance	0.55
Partial AIC	48.31
log-likelihood ratio test	0.23 on 1 df
-log2(p) of ll-ratio test	0.66

The results indicate that the herbicide has little impact on survival probability. The p-value is too high at 0.64 and the confidence intervals include 0 meaning the impact is not statistically significant. Plotting out the intervals makes his clear:



The intervals are almost centered around 0. There appears to be a slight bias indicating that Weedol may in fact improve survival probability but this is not credible and may be incidental and should therefore be ignored.

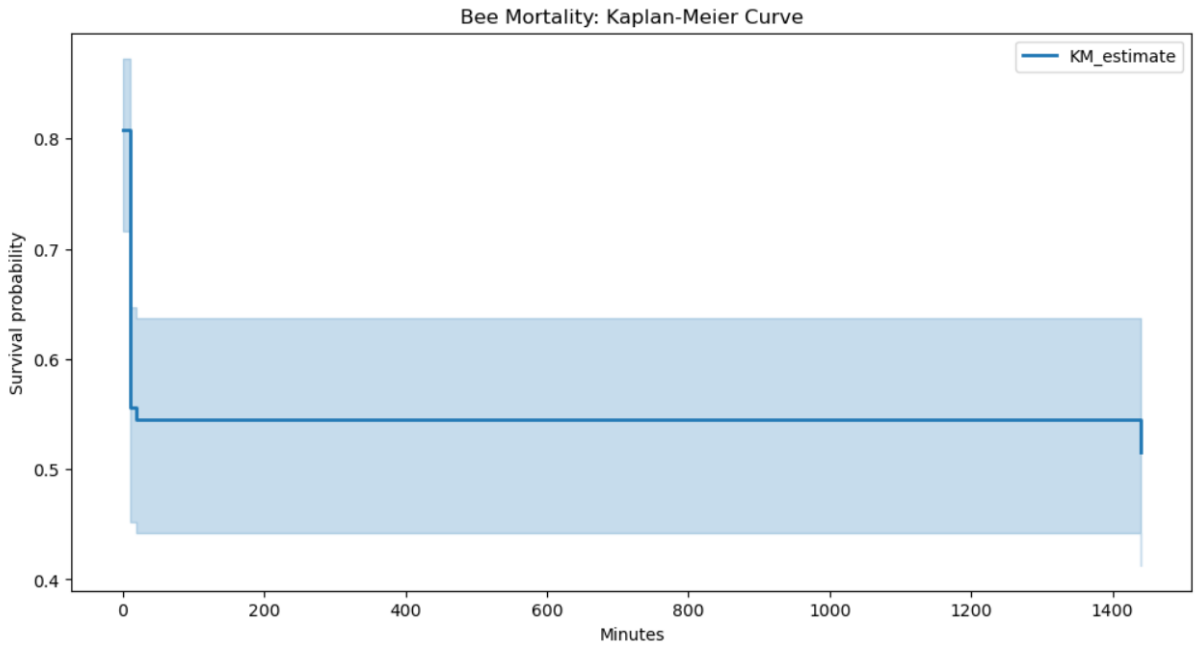
This is made clear when the effect of Weedol the herbicide on survival probability is noted.



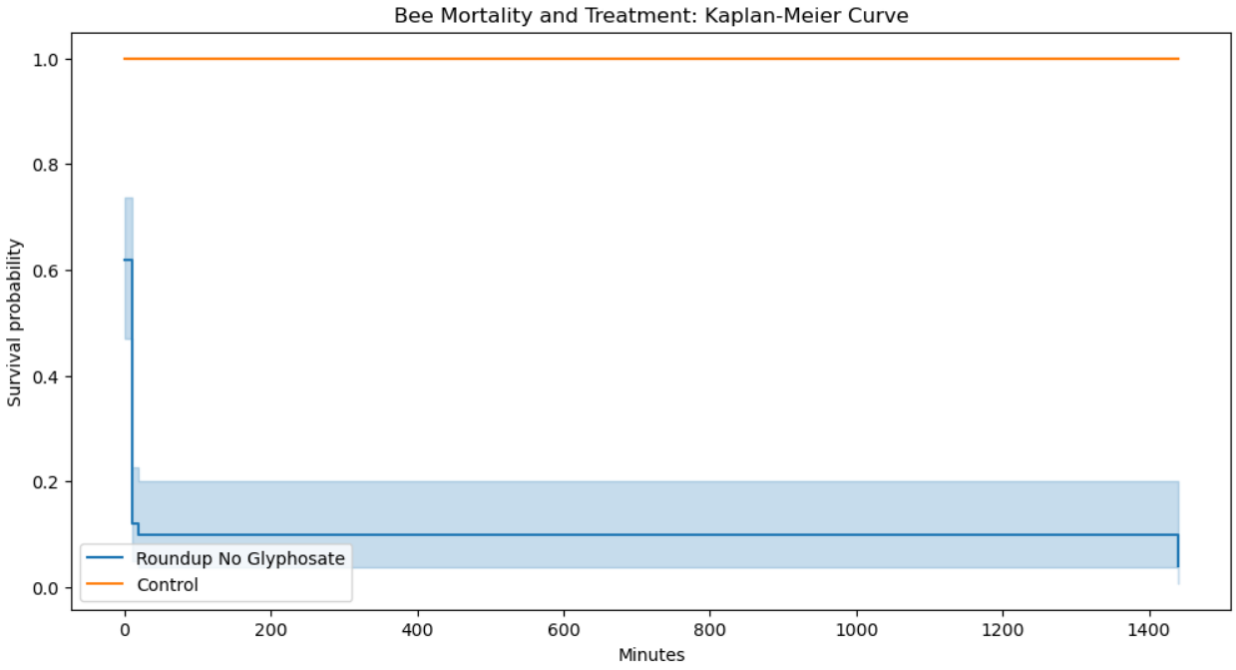
The plot shows nothing since there is no change on survival probability due to Weedol.

Dataset-3 Kaplan-Meier Curves:

Moving on, the next plot is of Kaplan-Meier curves for dataset 3:



The plot shows that something in the dataset clearly impacts survival probability significantly. By dividing the data according to treatment, this relationship is made obvious:



It is clear that the herbicide Roundup No Glyphosate is causing survival probability to fall dramatically.

Dataset-3 Cox Proportional Hazards Model:

The data is now to be modeled using Cox Proportional Hazards Model or Survival Regression. For this some feature engineering is done, one-hot encoding the Treatment column which is the feature column in the dataset.

```
df_u = df5[['Treatment','Event']]
df_d = pd.get_dummies(df_u, drop_first=True)
df_d['Time'] = df5.Time
df_d.head()
```

	Event	Treatment_Roundup No Glyphosate	Time
0	0.0	0	1440.0
1	0.0	0	1440.0
2	0.0	0	1440.0
3	0.0	0	1440.0
4	0.0	0	1440.0

Next, the data is fitted for Survival Regression and the results are noted.

```
cph = CoxPHFitter()
cph.fit(df_d, duration_col='Time', event_col='Event')
cph.print_summary()
```

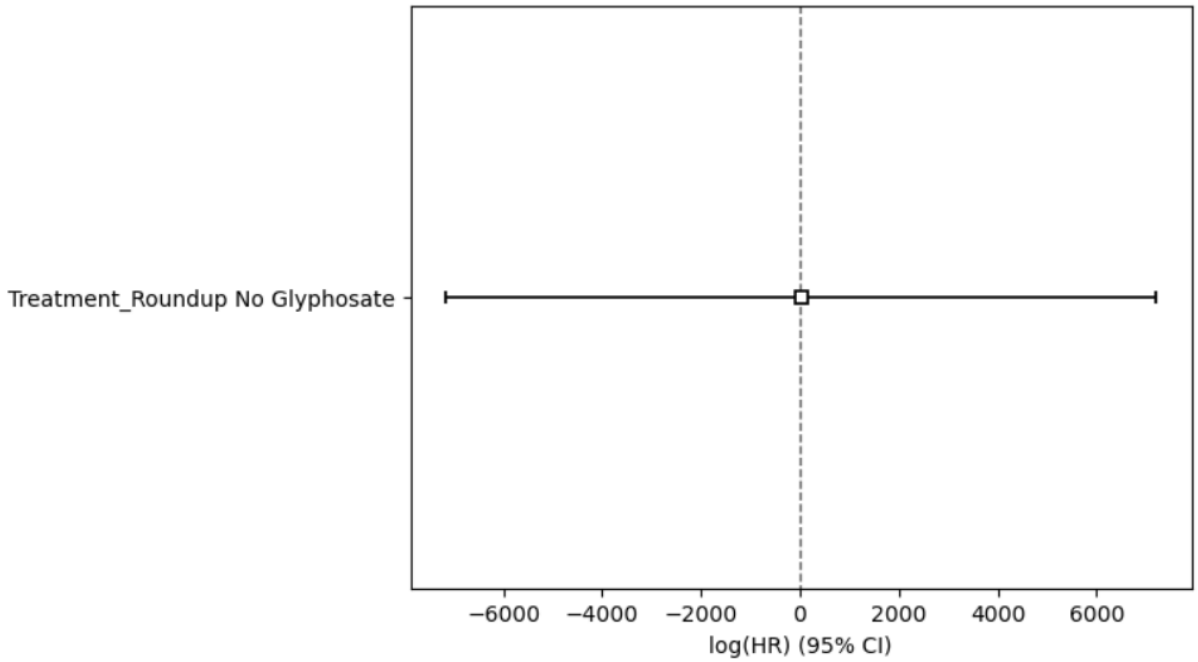
model	lifelines.CoxPHFitter
duration col	'Time'
event col	'Event'
baseline estimation	breslow
number of observations	99
number of events observed	48
partial log-likelihood	-147.78
time fit was run	2023-02-07 02:01:47 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Treatment_Roundup No Glyphosate	21.40	1.97e+09	3664.20	-7160.29	7203.10	0.00	inf	0.00	0.01	1.00	0.01

Concordance	0.88
Partial AIC	297.57
log-likelihood ratio test	117.88 on 1 df
-log2(p) of ll-ratio test	88.81

These results are confusing as it is evident from EDA and Kaplan-Meier curves plotted previously that the herbicide causes survival probability to drop. However, the regression results do not indicate this which is odd. The confidence intervals are massive

and p-value is 1. Something appeared to be causing an issue. Plotting the confidence intervals yielded this plot:



This is an unexpected result as the confidence intervals indicate that the result is statistically insignificant as they are centered at 0. It turned out that this was due to perfect separation in the dataset. All values of the Event column that were 1 occurred when Treatment_Roundup No Glyphosate column was 1.

```
df_c = df_d.loc[df_d['Event'] == 1]
a = df_c['Event'].sum()
print('Number of death events in dataset 3:', a)
b = df_c['Treatment_Roundup No Glyphosate'].sum()
print('Death events due to Roundup No Glyphosate:', b)
```

```
Number of death events in dataset 3: 48.0
Death events due to Roundup No Glyphosate: 48
```

To deal with this, a slight penalizer was added to the Survival Regression model. This penalizer works similar to Lasso and Ridge Regression. This resulted in the following model:

```
cph = CoxPHFitter(penalizer=0.001)
cph.fit(df_d, duration_col='Time', event_col='Event')
cph.print_summary()
```

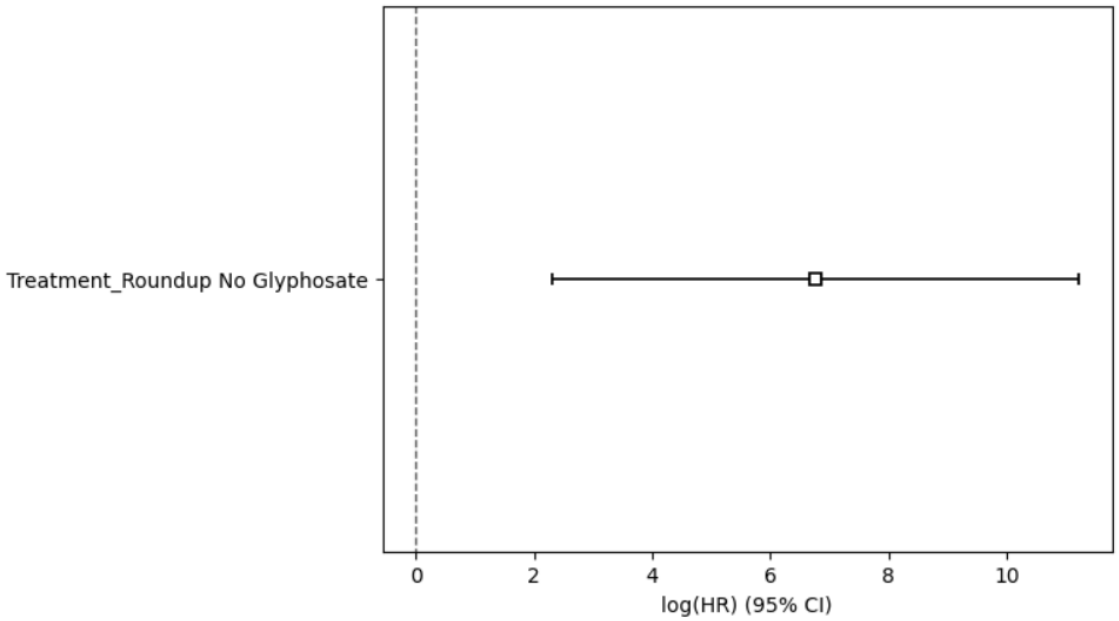
When this model was run, the results were as follows:

model	lifelines.CoxPHFitter
duration col	'Time'
event col	'Event'
penalizer	0.001
l1 ratio	0.0
baseline estimation	breslow
number of observations	99
number of events observed	48
partial log-likelihood	-148.53
time fit was run	2023-02-07 03:27:41 UTC

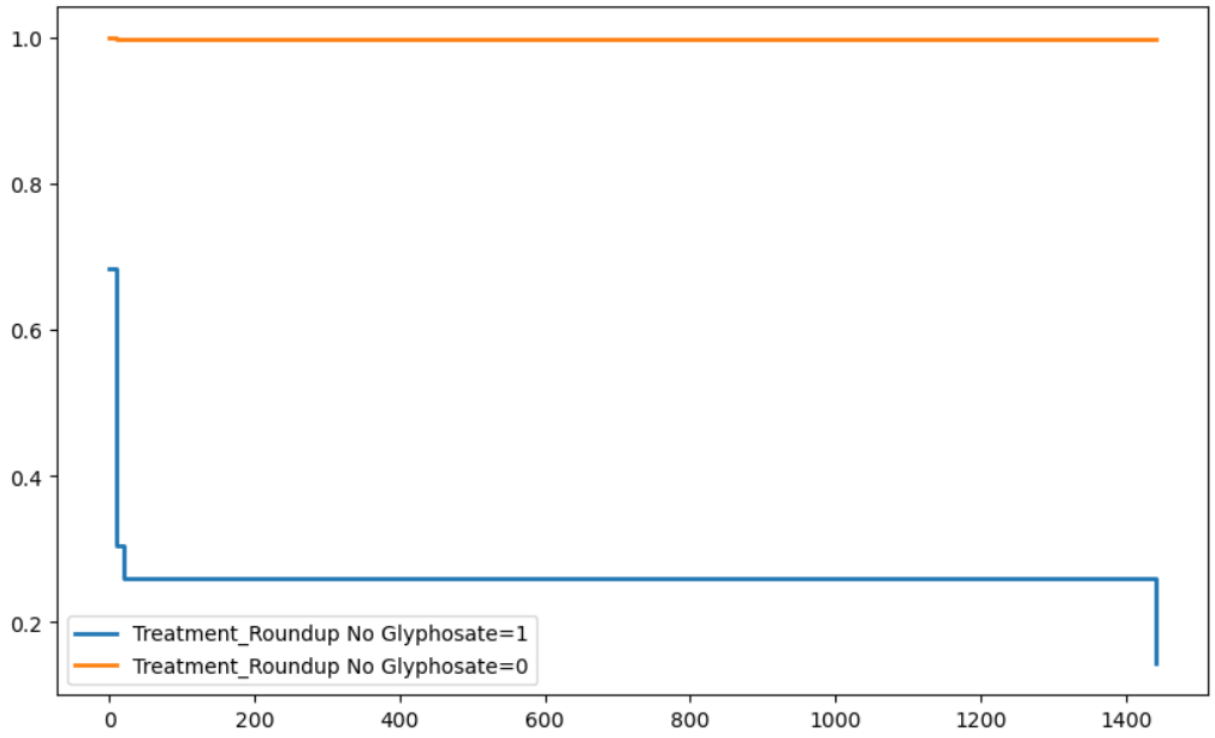
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
Treatment_Roundup No Glyphosate	6.76	863.33	2.28	2.30	11.22	9.95	74898.90	0.00	2.97	<0.005	8.39

Concordance	0.88
Partial AIC	299.05
log-likelihood ratio test	116.40 on 1 df
-log2(p) of ll-ratio test	87.73

As can be seen from the results, this led to rectification of the Perfect Separation problem. The p-value is now below 0.005 and confidence intervals do not include 0. Upon plotting, the following graph of confidence intervals is obtained:



Clearly, the result obtained is statistically significant as 0 is not included in the confidence interval. Furthermore, plotting survival probability with and without the herbicide further confirms this:



As can be seen rather clearly, using the herbicide Roundup No Glyphosate dramatically drops the survival probability of bees.

Discussion:

The survival analysis results for each dataset are important in understanding the impact of different herbicides. Roundup is one of the most popular brands of herbicide in the world today. Although regulatory authorities classify it as bee-safe, the experimental data used in this project proves otherwise.

The first dataset looks at two different brands of Roundup and their impact on bee survival. The analysis result from that survival analysis indicate that there is a positive correlation between the use of two Roundup products, Roundup Ready to Use 100% and Roundup ProActive.

The second dataset looks at another herbicide, Weedol and its impact on bee mortality. The results from that dataset indicate that Weedol does not have any effect on bee survival, reinforcing the view that there is something in Roundup products that reduces bee survival.

The third dataset looks at another Roundup product, Roundup No Glyosulphate. Glyosulphate is an active ingredient in some herbicides that is suspected to have

harmful effects on insects. However, Roundup No Glyosulphate turns out to have a very deleterious effect on bee survival as well, indicating that glyosulphate may not be the problem with Roundup products that causes bee mortality.

Key Findings:

The key findings from this analysis are as follows:

1. Not all herbicides are harmful to bees. Some such as Weedol, do not have any effect on bee mortality.
2. Roundup herbicides appear to have a very significant impact on bee mortality although the exact impact varies from product to product.
3. Glyosulphate an active ingredient in herbicides, thought to be responsible for declining bee populations is not the cause of bee mortality in case of Roundup products. Even without glyosulphate, some Roundup products still increase bee mortality.

Issues:

There were several issues with the analysis which need to be taken into account for future attempts to improve the results. These issues are:

1. The datasets used in this analysis were very small. If larger datasets were available, then the results may be more robust.
2. The data recorded in the datasets was not recorded at even time intervals but only at few select intervals. Even these varied from dataset to dataset. The results would be more robust if that was not the case.
3. Only a few herbicide products were used in the experiment. Making use of more brands may result in clearer results about Roundup and their competitors' products on bee populations.

Future Steps:

Some future steps for dealing with issues outlined above and more are listed below:

1. A more thorough analysis with several different herbicide brands and their effect on bumblebees.

2. Records of data from the experiment at even time intervals for more detailed insights.
3. A larger number of observations for the experimental data.
4. Similar experiments with many different breeds of bees for a more detailed analysis on impact of herbicides on bee populations.