STAT515 - Applied Statistics & Visualization for Analytics Final Project Report

Dataset : Genetically engineered (GE) corn, upland cotton and soybean varieties by State and United States, 2000-17

By Navya Shivaji Mote

Why this data set?

The agriculture industry has traditionally been supportive of technological advancement, particularly in the field of genetic crop improvement. For decades, the industry has been mixing naturally the genetic traits of seeds in the search for particularly robust varieties. The reaction of farmers to this new technology has been mixed. Some farmers have quickly adopted the technology. Other farmers, mindful of the controversy surrounding GE products, have hesitated to use GE seeds as part of their agricultural operations.

Farmers should understand both the benefits and concerns that are raised by the use of GE seeds. Benefits of the technology include increased crop yields, diminished use of pesticides and herbicides, and increased profits. Concerns that farmers should address before adopting the technology include the private contractual relations between farmers and seed companies, the environmental impacts of the technology, and the potential impacts of consumer concerns (both domestic and international) on the market for GE products.[1]

Questions I hope to answer using the data:

"The challenge of feeding the planet and doubling food supply in the next 36 years is the greatest challenge facing mankind today. … There are 7.2 billion people on the planet. There will be 9.6 billion by 2050. The demand for food will double… [Using GM food and data science is] the only thing that will enable us to feed the planet without encroaching on the forests and wetlands…. This represents a business opportunity, but from a societal perspective, it's very important." — Robert Fraley, CEO of Monsanto, Winner of the World Food Prize 2013.[2]

The area planted by GM crops has increased from 1.7 million hectares in 1996 to 179.7 million hectares (13% of world arable land[3]) in 2015[4]. Recently, state owned Chem-China made an acquisition of Syngenta (fourth largest seed company in the world and an active GM seed

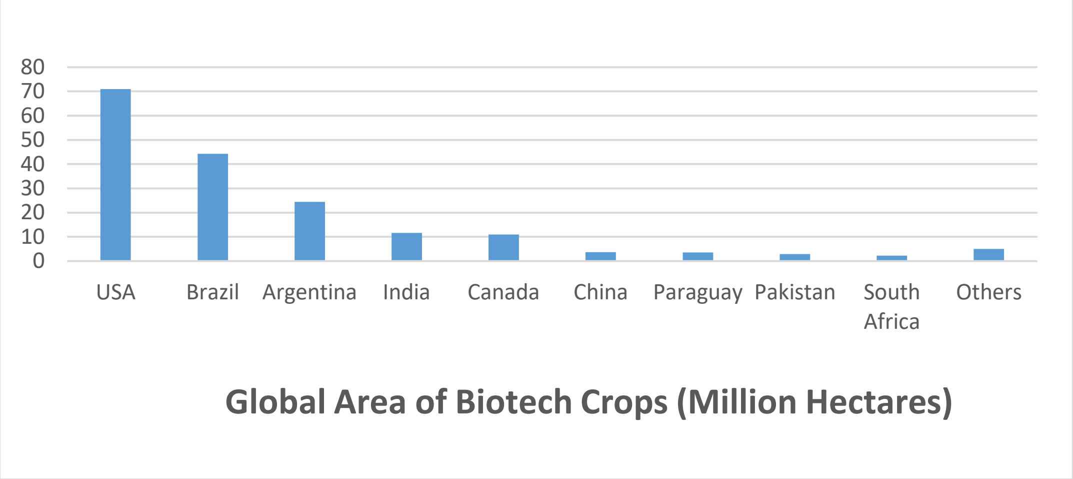
[1] Impacts of Genetically-Modified Crops and Seeds on Farmers
Prepared by David Kruft, Legal Research Assistant November 2001

[2] http://www.permaculturenews.org/files/Will_GM_Crops_Feed_the_World_cban_report_2014.pdf

[3] http://journal.georgetown.edu/genetically-modified-crops-and-their-role-in-international-development/

[4] http://www.isaaa.org/resources/publications/pocketk/16/default.asp

researcher) for $43 billion, making China a major investor in GM crop research. The graph below gives country-wise data for GM crop plantation area in 2015.



Source: http://www.isaaa.org/resources/publications/pocketk/16/default.asp

The crops planted in top five countries are given below:

Country	Biotech crops
USA	Squash, Soybean, Cotton, Maize, Canola, Sugar beet, Alfalfa, Papaya, Corn
India	Cotton
Argentina	Cotton, Maize, Soybean
Brazil	Maize, Cotton, Soybean
Canada	Soybean, Sugar beet, Canola, Maize

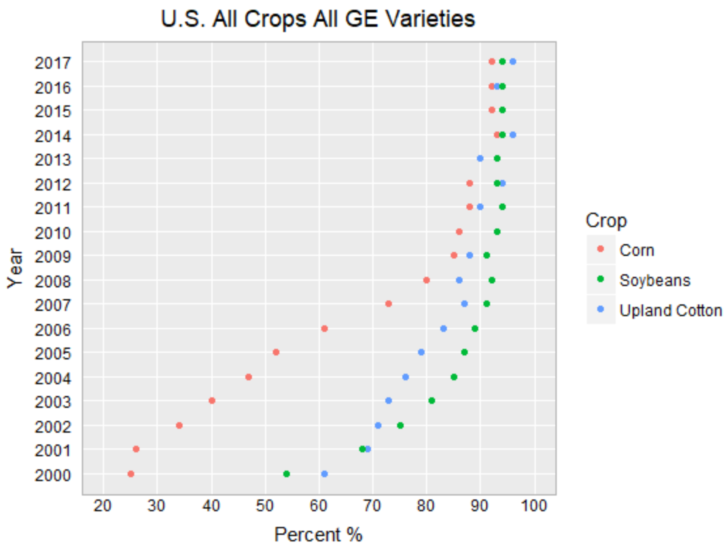
Source: http://www.isaaa.org

Since USA is the top producer of Genetically Engineered Crops in the world, I intend to show the progress of it through the years and across various states. This information is helpful to the USDA in looking back at data gathered over the years in order to make decisions. It could also be helpful for other countries to derive insightful information in order to apply relevant GE techniques.

Source of the data:

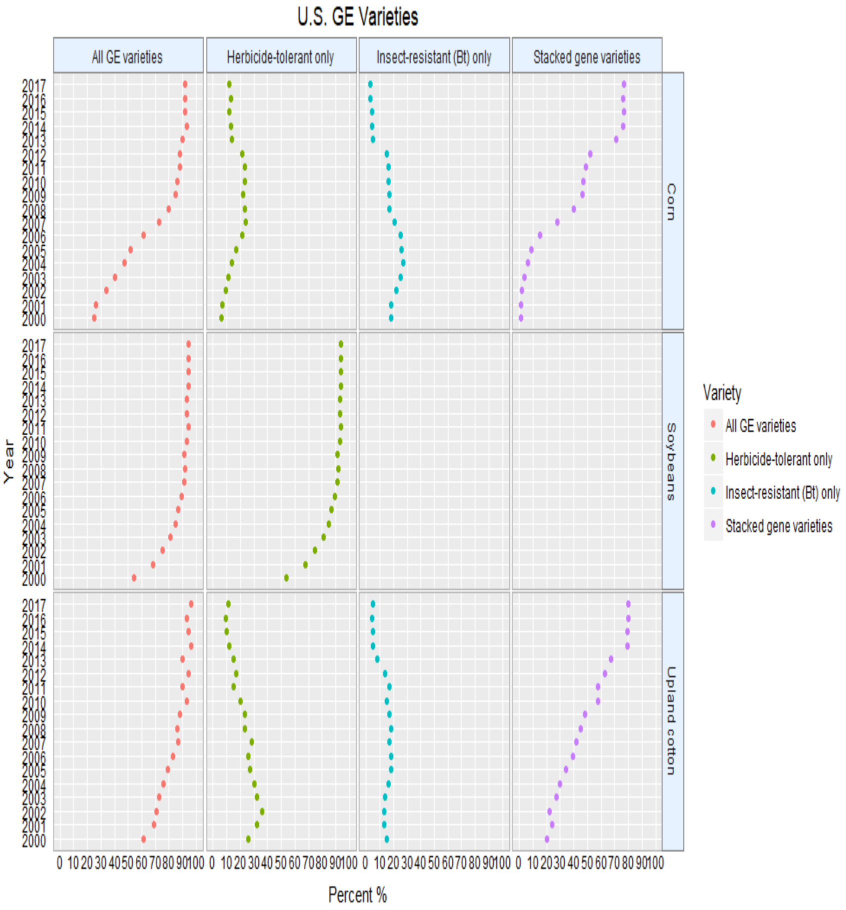
United States Department of Agriculture, Economic Research Service - Adoption of Genetically Engineered Crops in the U.S.

https://www.ers.usda.gov/data-products/adoption-of-genetically-engineered-crops-in-the-us/



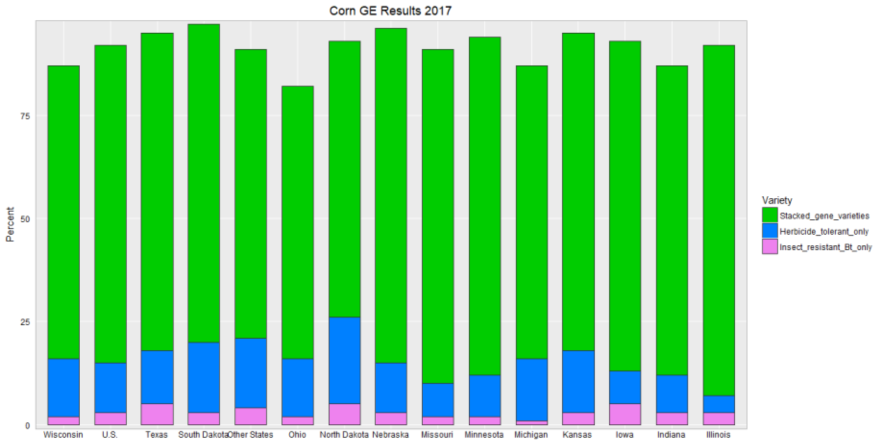
Graph 1

Using scatter plot in **Graph 1**, I am trying to show the increase in cultivation of Genetically Engineered corn, upland cotton and soybeans from 2000-2017 in the U.S. Looking at the above graph, we can see the gradual increase in the incorporation of GE crops through the years and also be able to compare them. We can say that the percentage of corn grown using GE in 2000 was about 20-30%, soybean was 50-60% and upland cotton was at 60%. Increasing gradually over the years, in 2017, the amount of corn, soybean and upland cotton grown using GE is 90-100%. As of now, almost all of the corn, soybean and upland cotton grown in the U.S. is genetically engineered.



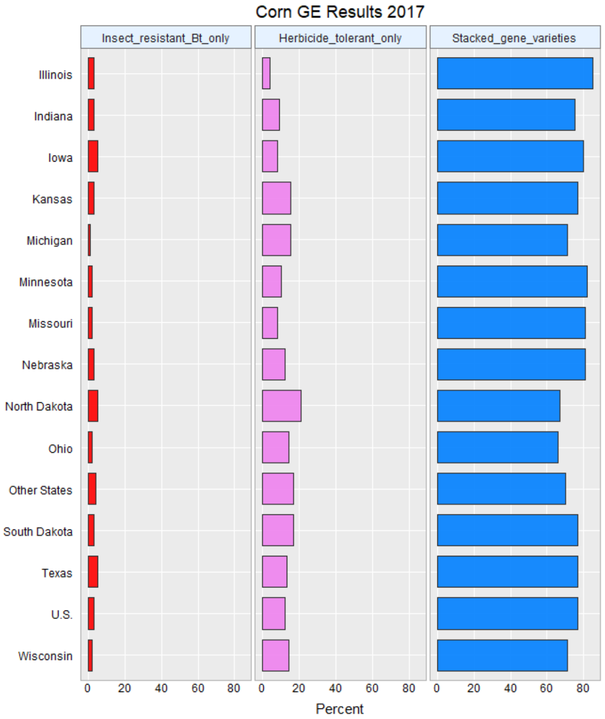
Graph 2

In **Graph 2** we are able to see three crops types, three genetically engineering varieties, and the results of all GE varieties put together, through a span of 17 years and the respective percentage value. I tried to visualize this graph using facet grid. With the help of this graph we can notice that the insect resistant(Bt) and the stacked gene variety is not at all used for growing soybeans. We can also observe that usage of insect resistant(Bt) and herbicide tolerant variety has reduced for corn and upland cotton since 2000.



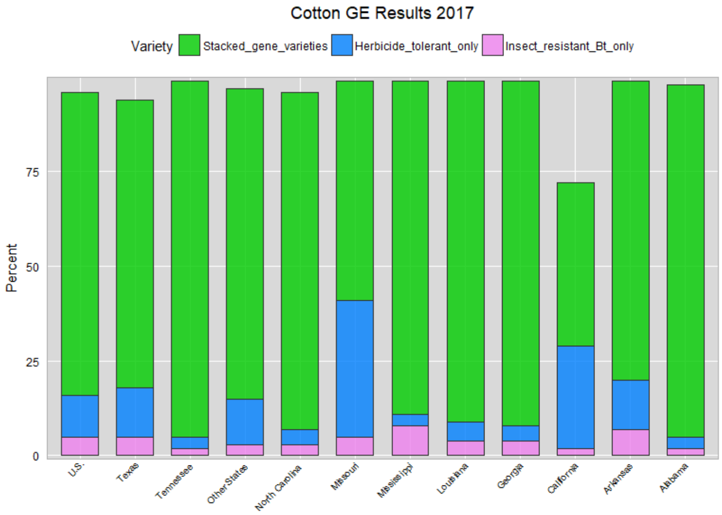
Graph 3

In **Graph 3**, I am trying to show the three GE varieties for corn using the bar plot. We can see that the most used technique in 2017 is stacked gene variety, second being the herbicide tolerant and the least being insect resistant(Bt). We can also see that these values are being shown for the states which are the major producers of GE corn.



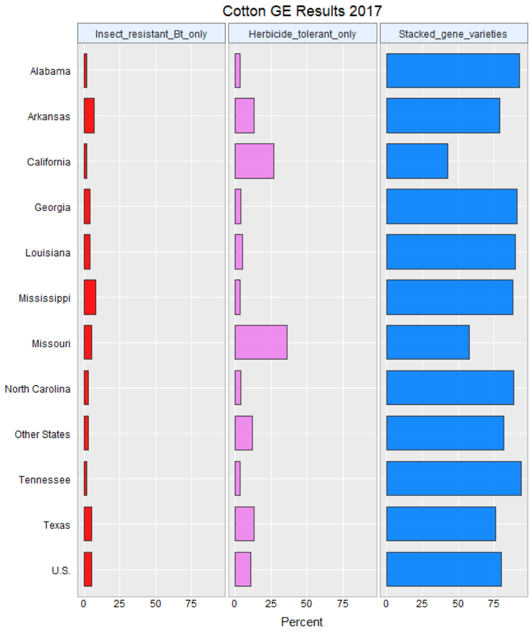
Graph 4

In **Graph 4**, I intend to show similar information as graph 3 while using the bar plot and facet grid. We can see the percentage of the respective GE technique used state-wise. The bar plot with facet grid comparatively gives us a more clear distinction between the percentage of the GE techniques used by each state. Looking at this graph we can derive that currently Iowa, North Dakota and Texas are the major producers of insect resistant(Bt) corn variety. North Dakota is also a major producer of the herbicide tolerant corn and Illinois is the largest producer of stacked gene corn variety.



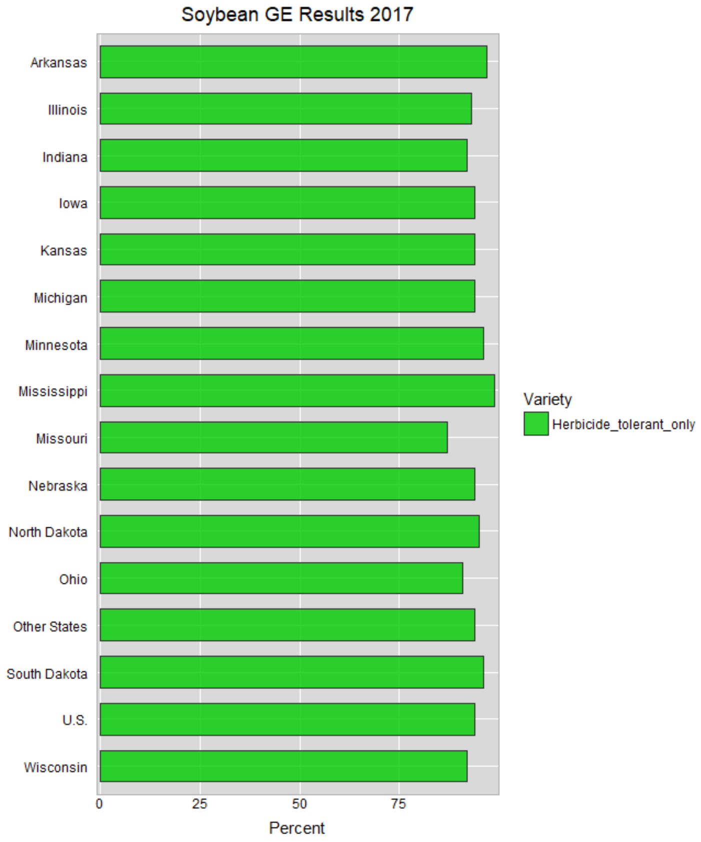
Graph 5

In **Graph 5**, I am showing the results of percentage cultivation of genetically engineered upland cotton in 2017 by the major state producers. I tried to do so while displaying the state name in slanting manner. The state names are more clearly visible here as they are not overlapping with each other.



Graph 6

Graph 6 is similar to graph 4 but the results displayed are for upland cotton in the year 2017. Looking at this graph we can derive that Mississippi and Arkansas are the largest producers of the insect resistant(Bt) variety. Missouri and California are the largest producers of the herbicide tolerant variety. Alabama and Tennessee are the largest producers of the stacked gene upland cotton variety.



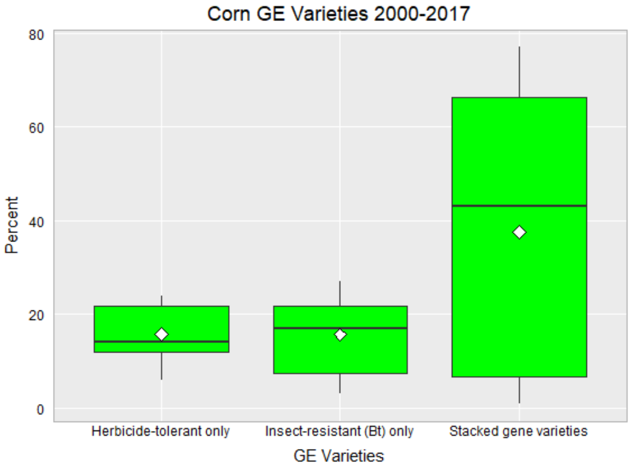
Graph 7

In **Graph 7** we can only see the herbicide tolerant variety as the soybean is majorly grown using this technique. Mississippi and Arkansas being the largest producer of soybeans using GE.

With the help of graphs 1-7 pictured above, I intend to show the dataset of genetically engineered crops and the techniques used in the U.S. We can observe that each crop has a technique that is more widely used than the others. We can also see the states which are the top producers of the genetically engineered crops.

With the help of these graphs we can also say that if a state is a major producer of a particular variety of crop, it does not mean that it is also a major producer of other crops of the same variety.

Overall, the usage of genetically engineered crops has increased in The United States over the years with stacked gene variety being the most popular for corn and upland cotton, and the herbicide tolerant for soybeans.

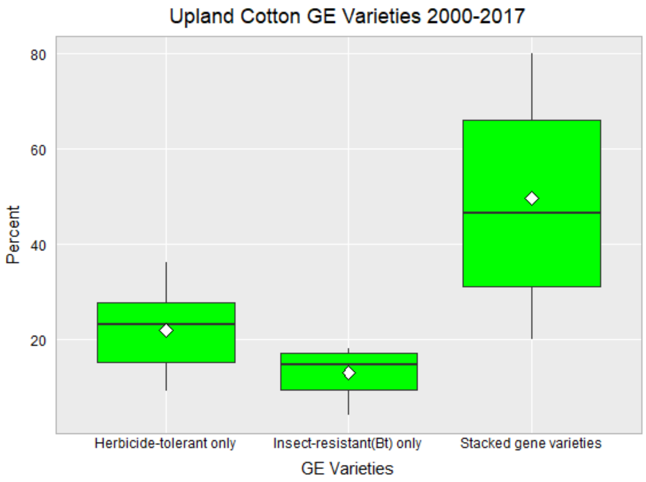


Graph 8

Graph 8 consists of data comprising all of the U.S. spanning from 2000-2017 for corn. The thick black lines are the medians. The ends of the green rectangle are the 1st and 3rd quartiles. We observe no outliers in this data. The vertical lines show the maximum and the minimum values of the distribution. The diamond represents the mean value.

We can observe that the median for herbicide tolerant variety is below the mean and the distribution is smaller when compared to the other techniques. The box plot for insect resistant(Bt) has a median above the mean. The box plot for stacked gene variety has the median way above the mean and is the largest distribution among the three techniques.

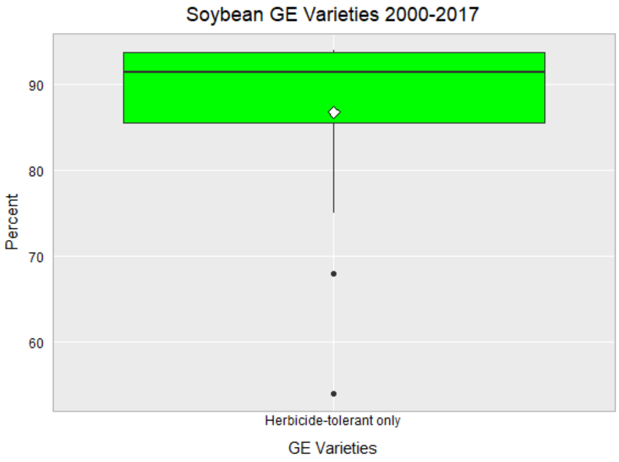
Looking at the distribution, we can analyze that the stacked gene variety is the most widely used technique, the insect resistant technique being second and the herbicide tolerant variety the least. Also, the percentage values for herbicide tolerant and insect resistant varieties has been consistent, neither increasing nor decreasing over the years. Whereas for stacked gene variety, the percentage values have been gradually increasing.



Graph 9

Graph 9 consists of data comprising all of the U.S. spanning from 2000-2017 for upland cotton. We can observe that the median for herbicide tolerant variety is above the mean and the distribution is larger than the insect resistant variety but smaller than the stacked gene variety. The median for insect resistant(Bt) is above the mean and the distribution is the smallest when compared to the other techniques. The median for stacked gene variety is below the mean and the distribution is the largest.

Looking at the graph, we can analyze that the stacked gene variety is the most widely used technique, second being the herbicide tolerant and at last is the insect resistant variety. Stacked gene variety has seen a gradual increase in the percentage values whereas herbicide tolerant and insect resistant variety does not show much variation.



Graph 10

Graph 10 consists of data comprising all of the U.S. spanning from 2000-2017 for soybean. For soybean, herbicide tolerant is the only widely used technique. We can observe that the median value is above the mean. There are 2 values that are the outliers in this box plot. The vertical line above the outliers is the location of the lower adjacent value. This the smallest data value inside the fence. Apart from the 2 outliers which are the years 2000 and 2001, the percentage values do not show much variation due to which we can understand that since 2002, the usage of herbicide tolerant variety has been in large consistently.

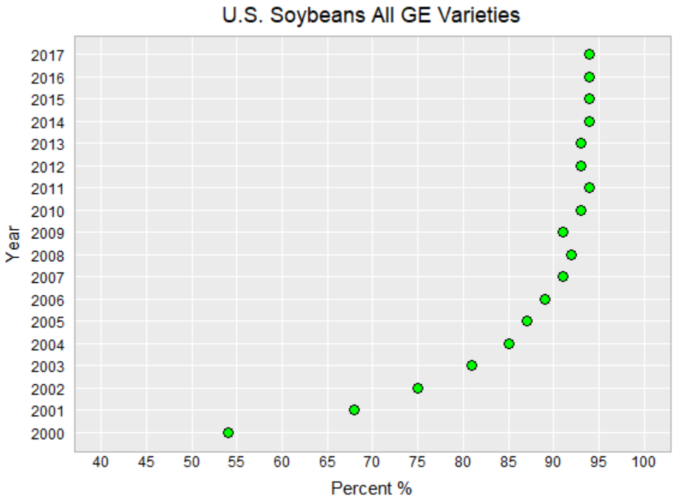
Conclusion:

After trying different visualization techniques, I feel that the graph 1 using scatter plot is helpful in giving us an overview of the data that we have. Graph 2, which is done using the facet grid is very helpful in observing the trend in the individual techniques and crops. It is with the help of this graph 2 that we are able to analyze that soybeans in this data has only one technique that is widely used. We can also say that for corn and cotton, stacked gene is the widely used technique over the years.

Looking at graphs 3 and 5, we can conclude that across all the states that are the major producers of corn and cotton mostly use the stacked gene variety. Also, the straight labelling of the states look neat when compared to the slant naming. But when space is a constraint, we can go with the slant labelling of x-axis to be able to accommodate more names. Graphs 4 and 6 help us look at the breakdown of the percentage for the respective states better.

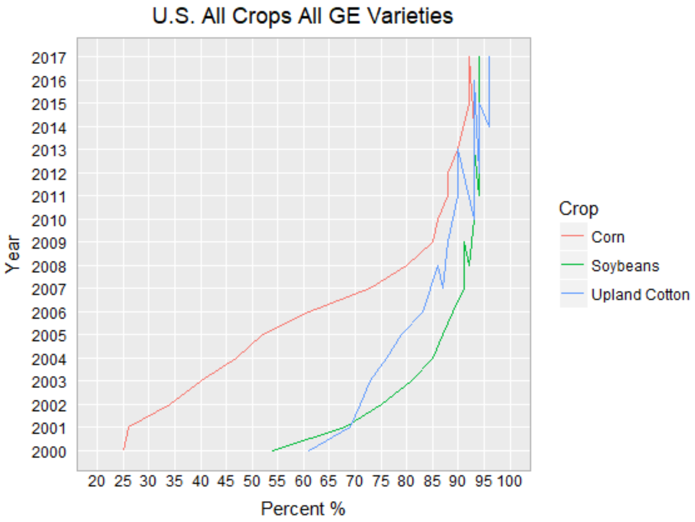
With respect to statistical analysis, I could observe the distribution using the box plot in graphs 8,9 and 10 for corn, upland cotton and soybeans respectively. We could observe the variation from 2000-2017. The mean, the median values and also the outliers.

APPENDIX



```
ggplot(Soybeans, aes( x=Percent, y=Year ) )+
  geom_point(shape=21, fill="green", size=3, color="black")+
  labs(x="Percent %",
       title="U.S. Soybeans All GE Varieties")+hw +
scale_x_continuous(breaks=seq(40,100,5),limits=c(40,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```

The above graph shows us the increase in GE usage for soybeans from 2000-2017 using dot plot

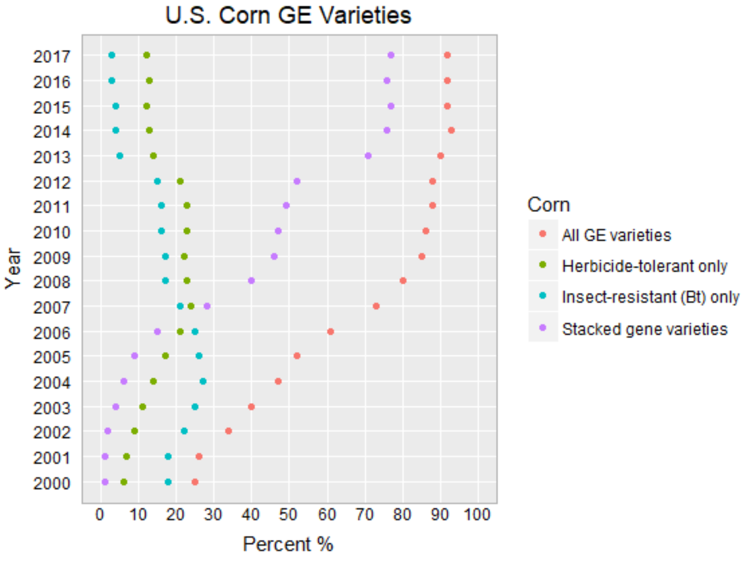


```
ggplot(Soybeans, aes(x=Percent,y=Year))+

geom_line(aes(colour = Crop))+

labs(x="Percent %",

title="U.S. All Crops All GE Varieties")+hw +

scale_x_continuous(breaks=seq(40,100,5),limits=c(40,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```

The above graph shows us the 3 crops and the overall GE usage using line plot. As the lines are overlapping, we are not able to see a clear distinction

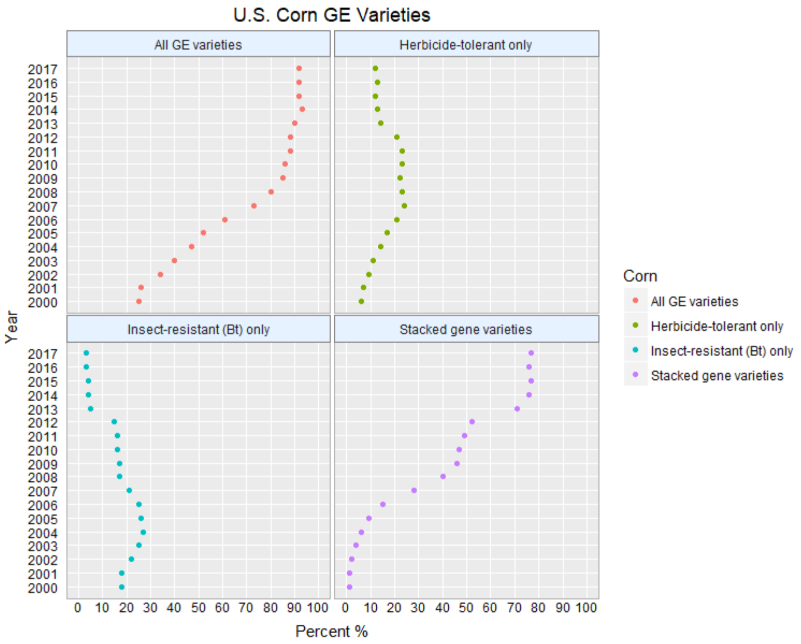
Using geomline() for comparison between the 3 crop types.

```
ggplot(Soybeans, aes(x=Percent,y=Year))+geom_line(aes(colour =
Crop))+ labs(x="Percent %",title="U.S. All Crops All GE
Varieties")+hw +
scale_x_continuous(breaks=seq(40,100,5),limits=c(40,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```



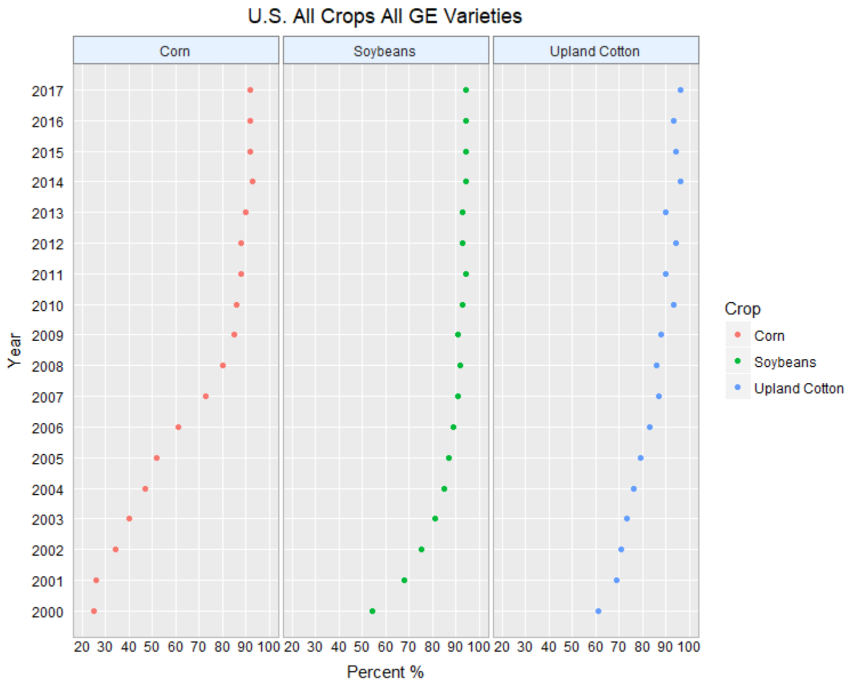
```
ggplot(corn,aes(x=Percent,y=Year))+

  geom_point(aes( color = Corn) ) +

  labs(x="Percent %",

       title="U.S. Corn GE Varieties")+hw +

  scale_x_continuous(breaks=seq(0,100,10),limits=c(0,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```

Above we can compare the different techniques used for corn from 2000-2017. Since we can see the comparison only for one crop, I did not feel that this would be a good graph to derive useful information.



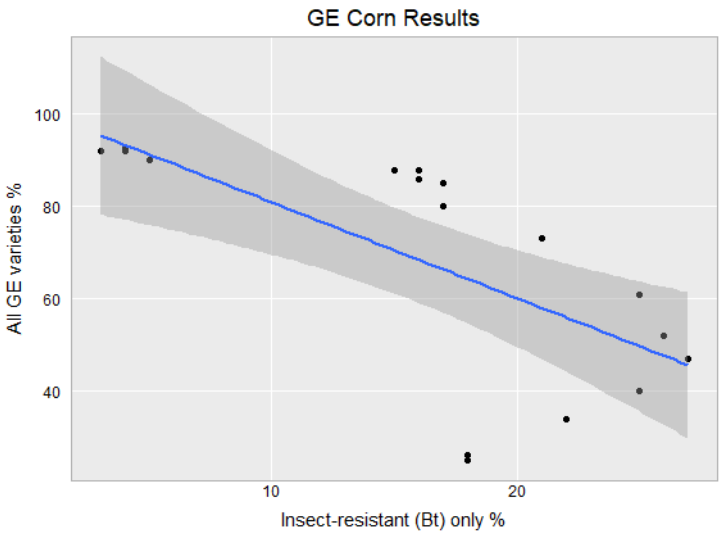
The above graph also shows techniques related to only one crop type. Therefore I did not feel that it would be useful to make a point.

```
ggplot(corn,aes(x=Percent,y=Year))+
  geom_point(aes( color = Corn) ) + facet_wrap(~Corn)+
  labs(x="Percent %",
       title="U.S. Corn GE Varieties")+hw +
  scale_x_continuous(breaks=seq(0,100,10),limits=c(0,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))


ggplot(data,aes(x=Percent,y=Year))+
  geom_point(aes( color = Variety) ) + facet_grid(Crop~Variety)+
  labs(x="Percent %",
       title="U.S. GE Varieties")+hw +
  scale_x_continuous(breaks=seq(0,100,10),limits=c(0,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```

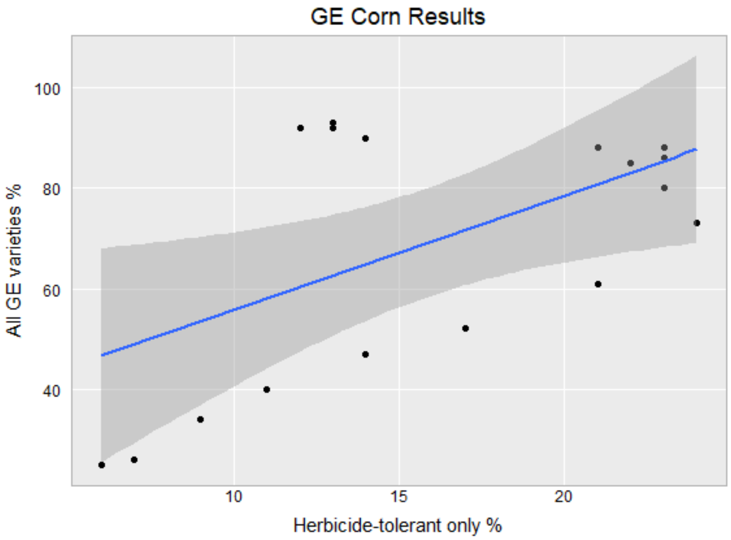


```
ggplot(Soybeans,aes(x=Percent,y=Year))+
  geom_point(aes( color = Crop) ) + facet_wrap(~Crop)+
  labs(x="Percent %",
       title="U.S. All Crops All GE Varieties")+hw +
  scale_x_continuous(breaks=seq(20,100,10),limits=c(20,100)) +
scale_y_continuous(breaks=seq(2000,2017,1),limits=c(2000,2017))
```

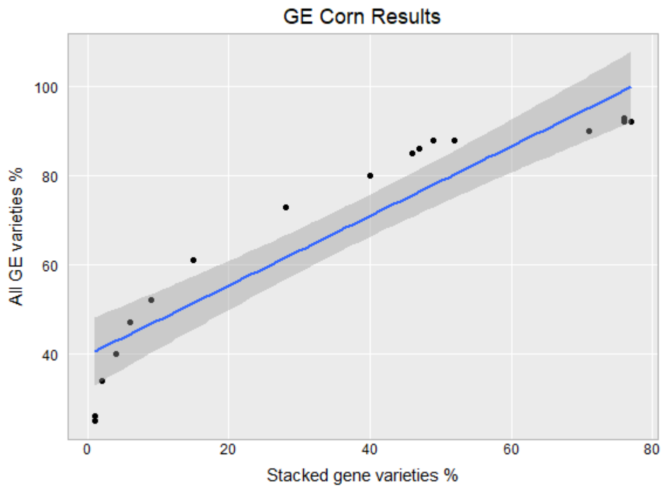
The above graph shows the overall GE performance for the 3 crop types. Since all 3 of them are gradually increasing, I did not find this graph very interesting.



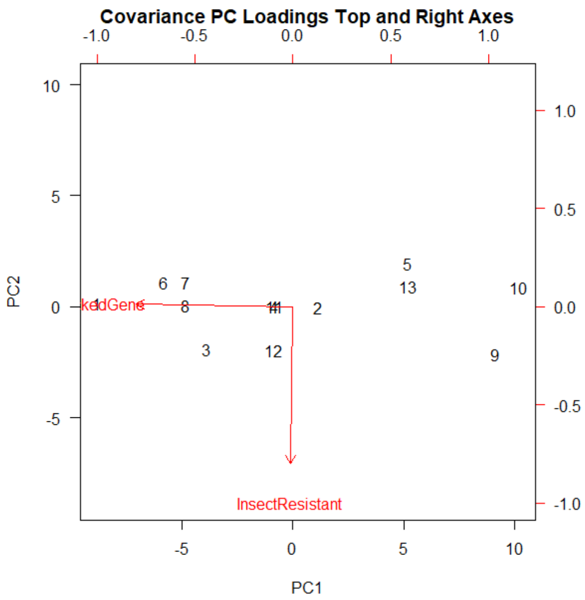
I tried linear regression to compare the performance of all GE varieties to the insect resistant variety for corn. But the comparison does not seem to yield useful information.

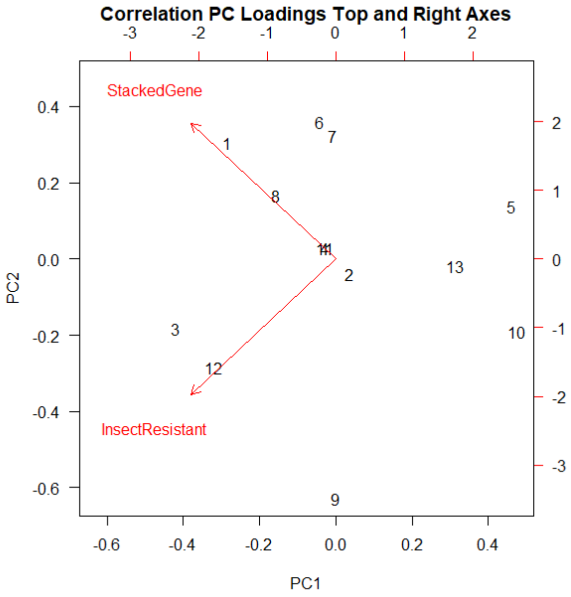


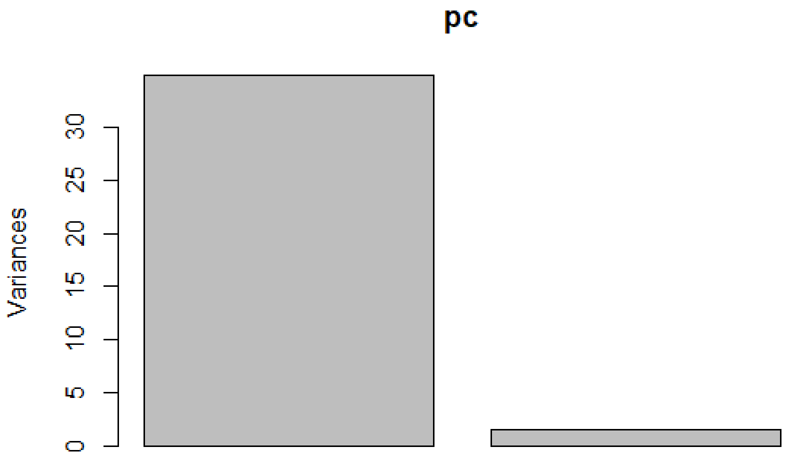
I tried linear regression to compare the performance of all GE varieties to the herbicide tolerant variety for corn. But the comparison does not seem to yield useful information.

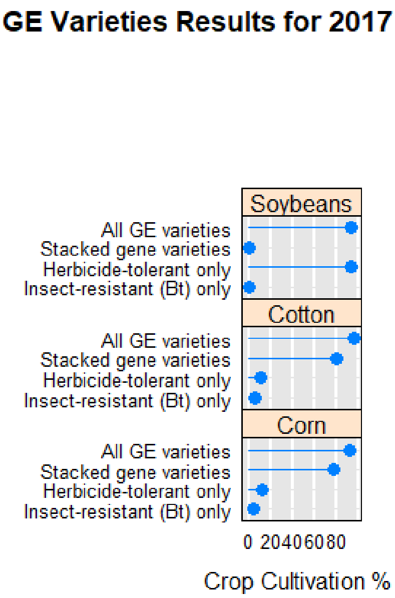


I tried linear regression to compare the performance of all GE varieties to the stacked gene variety for corn. But the comparison does not seem to yield useful information.

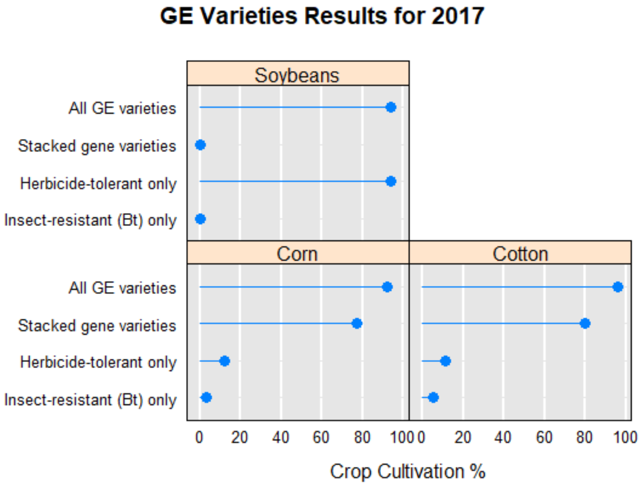



Correlation PC Loadings Top and Right Axes
-3
-2
-1
0
1
2
StackedGene
InsectResistant
1
2
3
5
6
7
8
9
10
11
12
13
0.4
0.2
0.0
-0.2
-0.4
-0.6
PC2
2
1
0
-1
-2
-3
-0.6
-0.4
-0.2
0.0
0.2
0.4
PC1

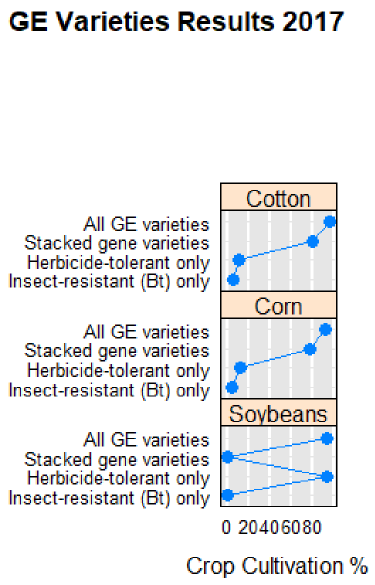

pc
Variances
0
5
10
15
20
25
30



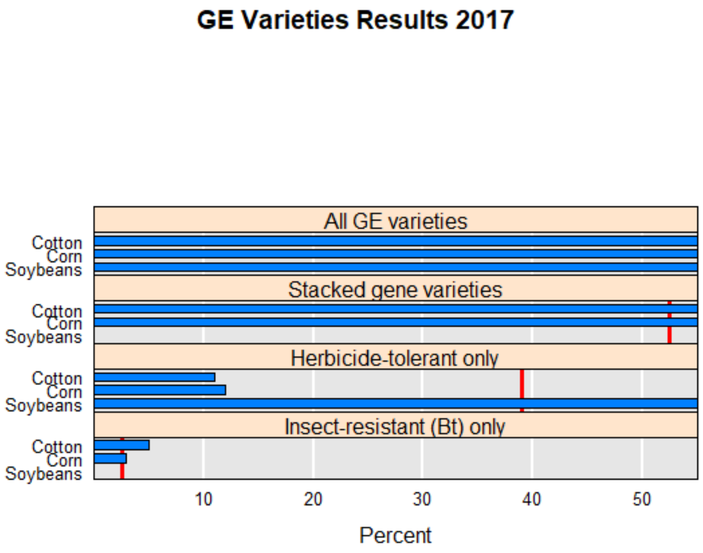
In the above graph we can see that for soybeans, herbicide variety is the most popular. For cotton it is stacked gene and even for corn. But we cannot see the results with respect to the years. Therefore, I did not feel that this graph would be a good fit for my dataset.



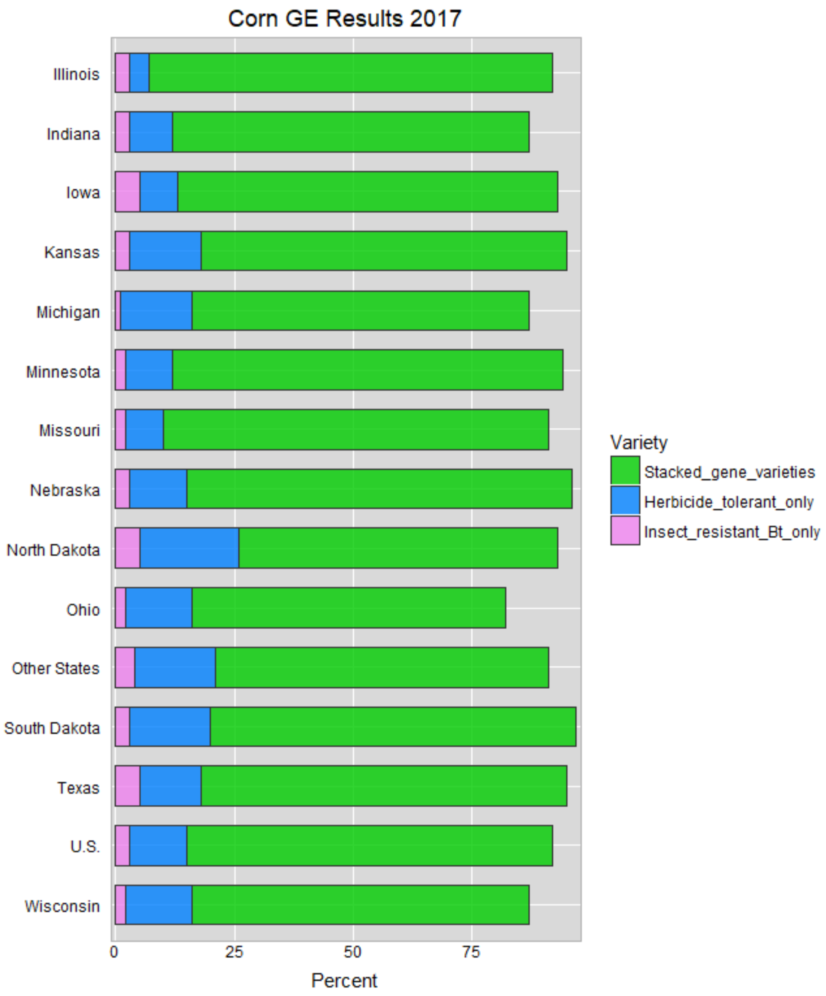
Since the results for corn and cotton look almost the same, I did not find this graph interesting enough.



This graph is confusing with respect to my dataset. I did not feel that it could be useful



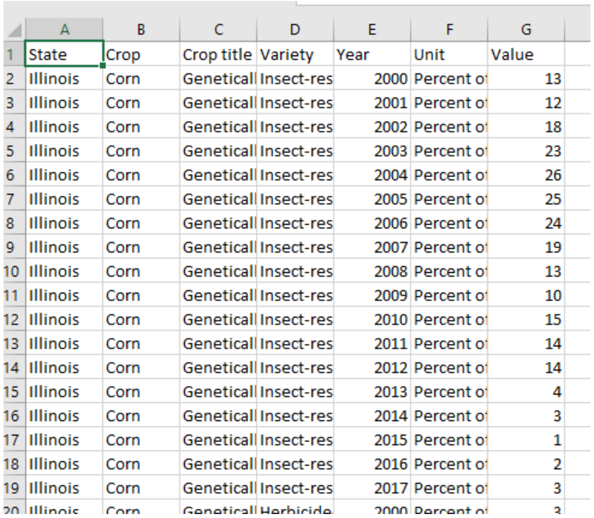
With this graph I did not feel I could derive useful information for my dataset as it tries to compare each technique across the crops. While soybeans does not use any other GE technique other than herbicide tolerant, the results are quite generic



The above graph is a horizontal bar plot that I tried. I chose to use the vertical bar plot in my project as I was already using the horizontal bar plot with facet grid. I wanted to show some variety.

Effort:

The dataset has 7 columns – State, Crops, Crop title, Variety, Year, Unit and Value.



	A	B	C	D	E	F	G
1	State	Crop	Crop title	Variety	Year	Unit	Value
2	Illinois	Corn	Geneticall	Insect-res	2000	Percent o	13
3	Illinois	Corn	Geneticall	Insect-res	2001	Percent o	12
4	Illinois	Corn	Geneticall	Insect-res	2002	Percent o	18
5	Illinois	Corn	Geneticall	Insect-res	2003	Percent o	23
6	Illinois	Corn	Geneticall	Insect-res	2004	Percent o	26
7	Illinois	Corn	Geneticall	Insect-res	2005	Percent o	25
8	Illinois	Corn	Geneticall	Insect-res	2006	Percent o	24
9	Illinois	Corn	Geneticall	Insect-res	2007	Percent o	19
10	Illinois	Corn	Geneticall	Insect-res	2008	Percent o	13
11	Illinois	Corn	Geneticall	Insect-res	2009	Percent o	10
12	Illinois	Corn	Geneticall	Insect-res	2010	Percent o	15
13	Illinois	Corn	Geneticall	Insect-res	2011	Percent o	14
14	Illinois	Corn	Geneticall	Insect-res	2012	Percent o	14
15	Illinois	Corn	Geneticall	Insect-res	2013	Percent o	4
16	Illinois	Corn	Geneticall	Insect-res	2014	Percent o	3
17	Illinois	Corn	Geneticall	Insect-res	2015	Percent o	1
18	Illinois	Corn	Geneticall	Insect-res	2016	Percent o	2
19	Illinois	Corn	Geneticall	Insect-res	2017	Percent o	3
20	Illinois	Corn	Geneticall	Herbicide	2000	Percent o	3

In order to represent the data using different techniques, I had to retrieve the relevant rows and columns. As it was a dataset that was new to me, there were a couple of errors while trying out the techniques which I rectified on my own. I tried out each and every technique taught in class and used the ones that depicted my data the best in my project. I also tried the regression models but as my data did not have many attributes, I selected the box plot distribution instead. The dataset that I chose spans 17 years and has a continuous variable which is the percentage. I also tried time series technique but for that I would need monthly data as well. The challenge was to represent 17 years of data, while knowing that the usage of GE has been increasing over the years and yet to be able to derive useful information out of it.