INTRODUCTION:

In 2006, American honey agriculture has faced a dramatic reduction in the honeybee population. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Twelve years later, some industries are observing recovery, but the American honey industry is still largely struggling. With this, US has started importing honey overseas. This project provides a perspective to the producer about how the yield production and its value changed over the years from 1998 to 2016. In this project, the focus is on predicting honey production. This provides an idea to decision-makers whether to invest in honey production or not.

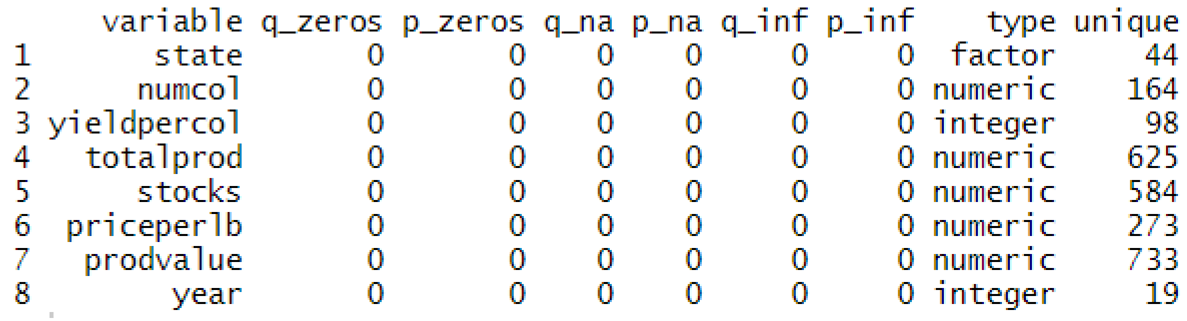
PROBLEM DEFINITION AND FORMULATION:

The dataset has eight features including Year (1998-2016) stored as both numeric and factor data types including 795 entries.

Attributes	Description
numcol	Number of honey producing colonies.
yieldpercol	Honey yield per colony. Unit is pounds
totalprod	Total production (numcol x yieldpercol). Unit is pounds
stocks	Refers to stocks held by producers. Unit is pounds
priceperlb	Refers to average price per pound based on expanded sales. Unit is dollars.
prodvalue	Value of production (totalprod x priceperlb). Unit is dollars.

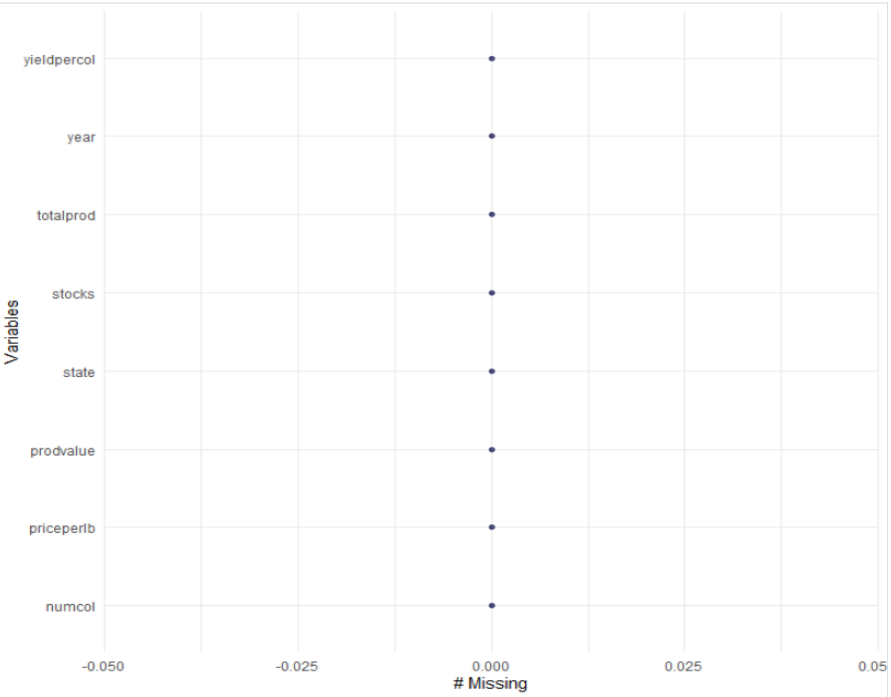
Exploratory Data Analysis & Data Preparation:

Initial Exploratory Data Analysis has been performed to know the structure of the data in both visual and statistical manner. The statistical table below gives the list of variables, the number of missing values and the percentage of missingness along with the data types. The column which represents missingness has all zeros that mean the dataset is free from NA values. The visual representation of checking missing values can also be found below.



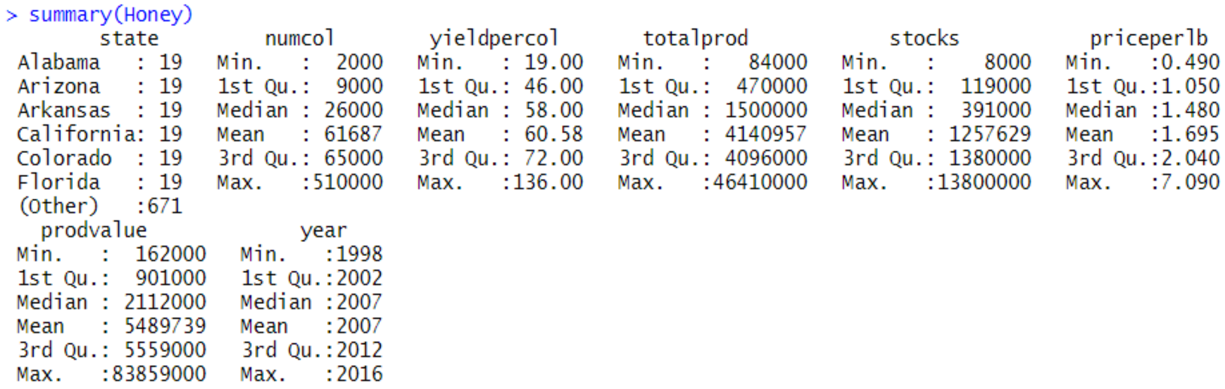
```
     variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
1       state       0       0    0    0     0     0  factor     44
2      numcol       0       0    0    0     0     0 numeric    164
3 yieldpercol       0       0    0    0     0     0 integer     98
4   totalprod       0       0    0    0     0     0 numeric    625
5      stocks       0       0    0    0     0     0 numeric    584
6  priceperlb       0       0    0    0     0     0 numeric    273
7   prodvalue       0       0    0    0     0     0 numeric    733
8        year       0       0    0    0     0     0 integer     19
```

Missingness:



Summary:

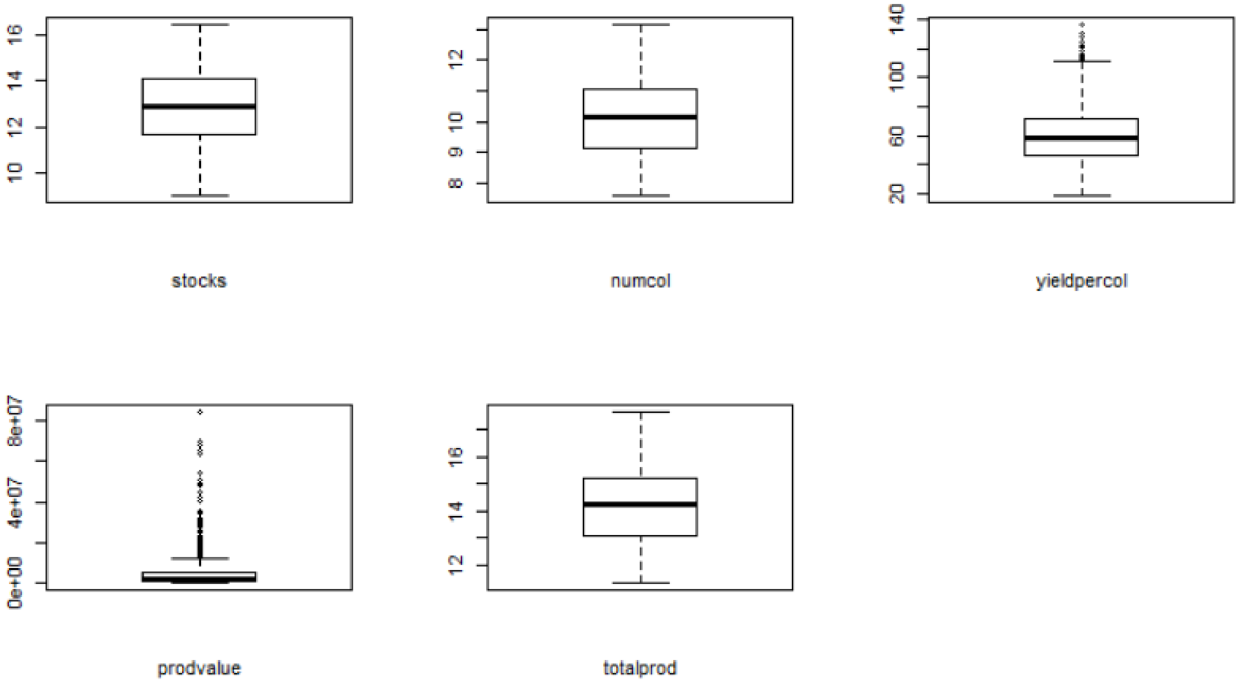
The summarization of the attributes can be found using summary method that provides various statistical results such as min, max, Inter Quartile Range (1st and 3rd). For instance, the value of the “prodval” attribute ranges from 162000 to 83859000. This statistical table is quite useful in understanding the distribution of the attributes.



```
> summary(Honey)
        state           numcol          yieldpercol        totalprod            stocks            priceperlb
 Alabama   : 19   Min.   :  2000   Min.   : 19.00   Min.   :   84000   Min.   :    8000   Min.   :0.490
 Arizona   : 19   1st Qu.:  9000   1st Qu.: 46.00   1st Qu.:  470000   1st Qu.:  119000   1st Qu.:1.050
 Arkansas  : 19   Median : 26000   Median : 58.00   Median : 1500000   Median :  391000   Median :1.480
 California: 19   Mean   : 61687   Mean   : 60.58   Mean   : 4140957   Mean   : 1257629   Mean   :1.695
 Colorado  : 19   3rd Qu.: 65000   3rd Qu.: 72.00   3rd Qu.: 4096000   3rd Qu.: 1380000   3rd Qu.:2.040
 Florida   : 19   Max.   :510000   Max.   :136.00   Max.   :46410000   Max.   :13800000   Max.   :7.090
 (Other)   :671
   prodvalue             year     
 Min.   :  162000   Min.   :1998  
 1st Qu.:  901000   1st Qu.:2002  
 Median : 2112000   Median :2007  
 Mean   : 5489739   Mean   :2007  
 3rd Qu.: 5559000   3rd Qu.:2012  
 Max.   :83859000   Max.   :2016  
```

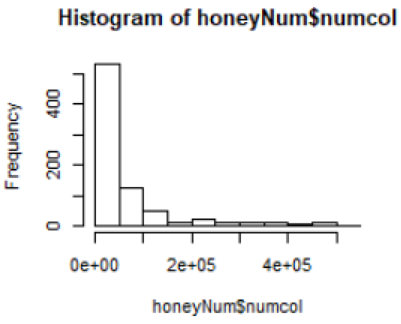
Outliers:

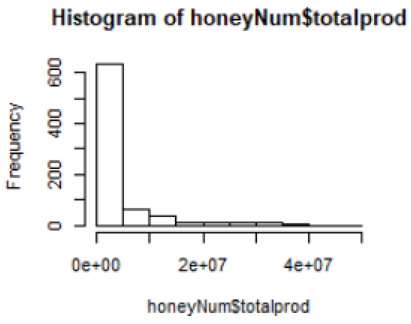
The presence of outliers in a dataset may dangerously affect the model. Therefore, outlier detection and treating them are crucial steps to take care of. Honey dataset indeed has outliers especially in “prodvalue”. The yieldpercol attribute has also outliers. Since it is not always a good idea to remove outliers and the present of the proportion of the outliers is considerably high, they all are taken to train the model. Because the given dataset is small and treating outliers may change the structure of the dataset, the outliers have not been deleted.

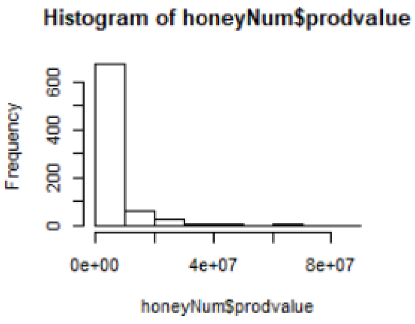


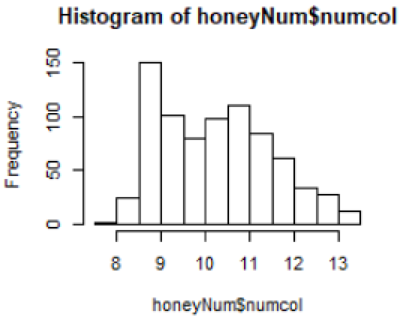
<u>Histograms to check the Distribution:</u>

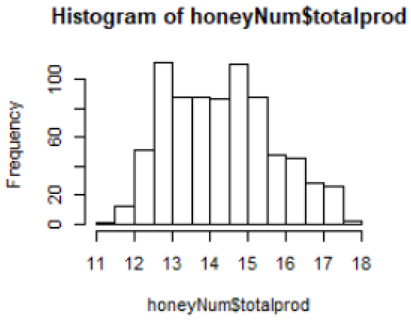
Below plots represent histograms of specific numeric variables that show the distribution of data among each variable. The plots from the first row are the original distribution of variables such as numcol, totalprod, prodvalue. The distributions of all variables are right-skewed. To avoid unusual results while predicting the data, the values have been transformed logarithmically and now the distribution looks normal which can be found in the second row in the below plot.

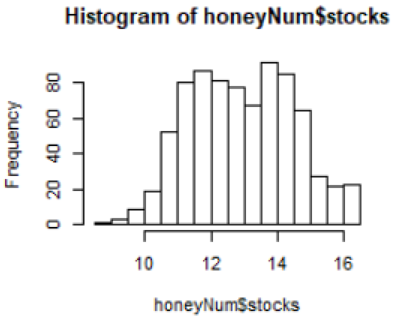






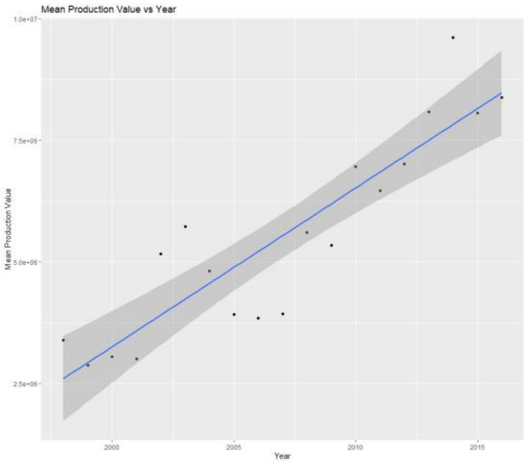


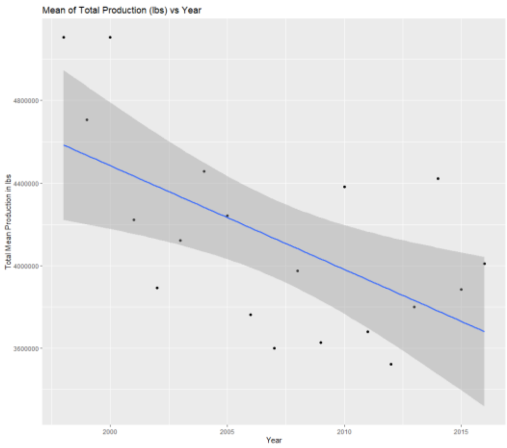




Insights:

Two scatter plots below with regression lines enable us to understand the positive and the negative the trend between production value and year as well as totalpro and year.

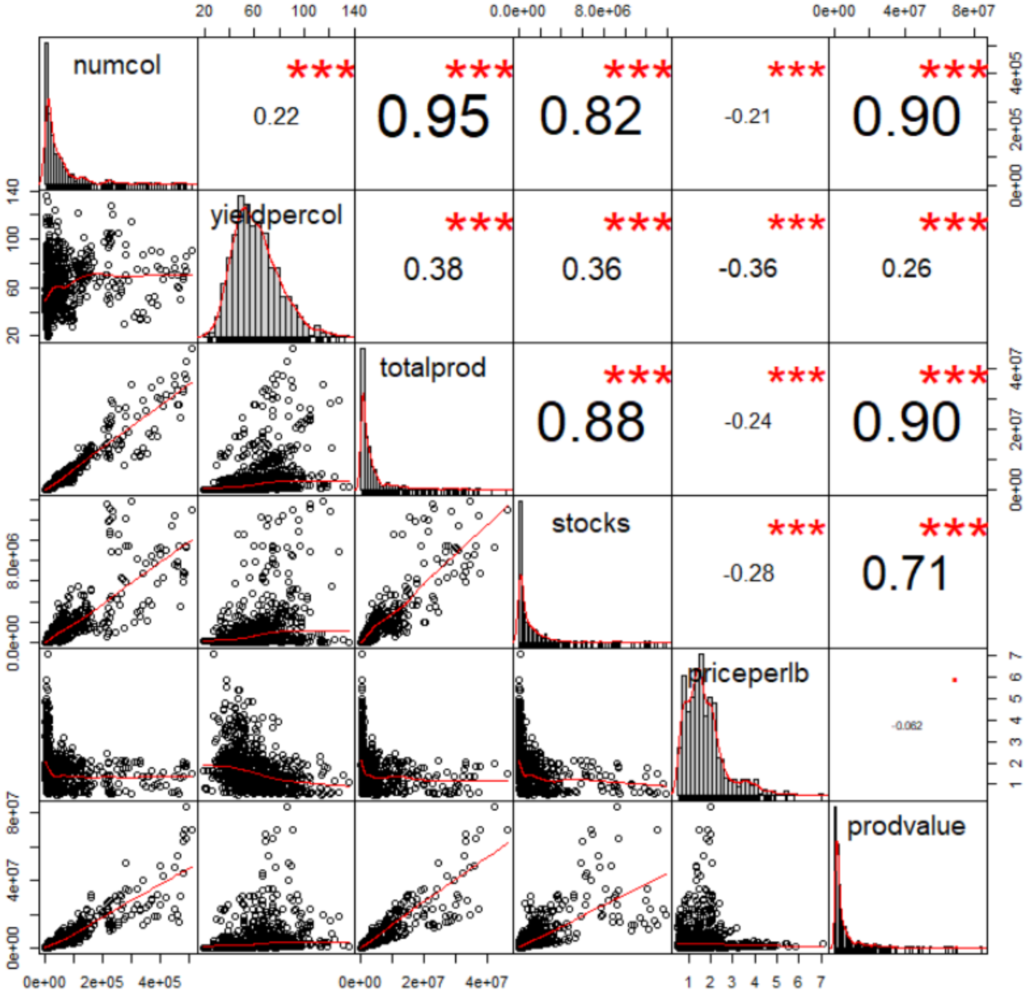




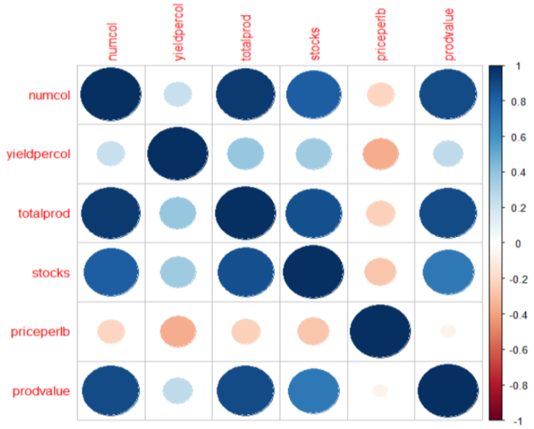
Correlation Analysis:

The correlation chart gives quite informative details. The main statistics such as correlation, linear regression, data distribution details can be found in one plot. From the chart, the combinations (numcol, totalprod), (numcol, prodval), (totalprod, approval) having correlation 95%, 90%, and 90% respectively. Almost all variables have either a strong positive or negative correlation among themselves.

The correlation matrix is given below also provides the same information as the correlation chart does. Though there is not much information available, it always helps to have a quick glance at how the variables are related to each other.



Correlation Matrix:



Solutions and Discussion:

To predict the “totalprod” values, created the fitted model using a multiple linear regression algorithm in the RStudio platform. Testing the predicted model on the test data has yielded almost accurate performance with a 0.002 error rate. Used ‘RMSE’ to evaluate performance.

RMSE:

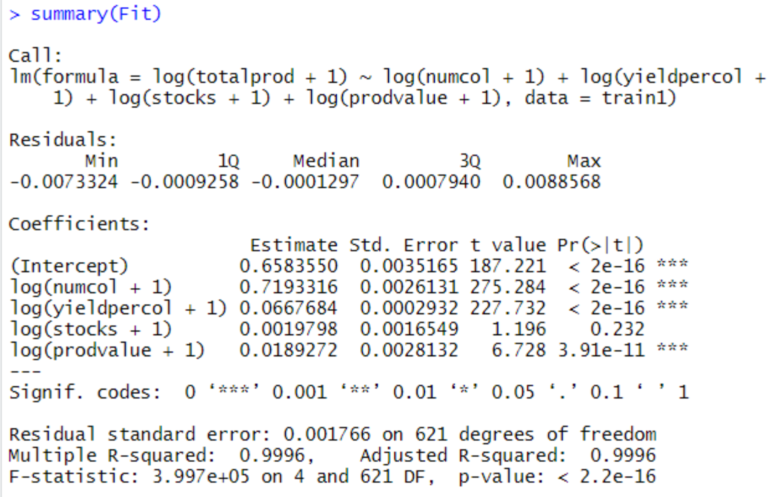
Root Mean Squared Error is one of the performance evaluation metrics which calculates the performance by taking the predicted and actual values. The equation to calculate RMSE is given below. From the equation, “n” is the number of observations and $\hat{y}_i - y_i$ is the difference between the actual and predicted values.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

RMSE Value: 0.00204894

Summary of the Fit:

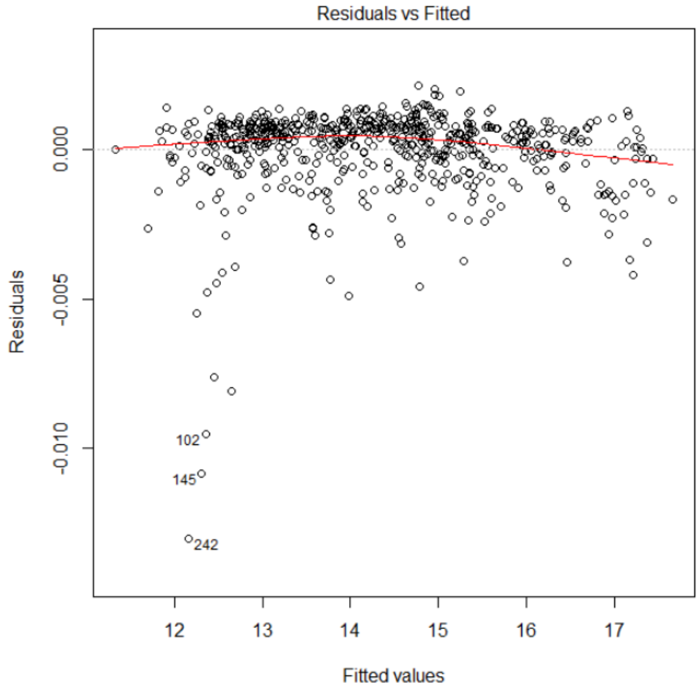
The summary of the predicted model gives the coefficients of each used attribute while building the model, t-value, p-value, R-Squared Value, etc. Four variables numcol, yield, prodvalue have a high correlation with the predictor variable. The coefficients of the considered variables with respect to predictor have been estimated along with the standard error. It also provides the significance levels. Therefore, we can consider the model for prediction. Finally, considering the p-value which is negligible, we can conclude that considering built model yields good prediction values with low error rate.



```
> summary(Fit)

Call:
lm(formula = log(totalprod + 1) ~ log(numcol + 1) + log(yieldpercol + 
    1) + log(stocks + 1) + log(prodvalue + 1), data = train1)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0073324 -0.0009258 -0.0001297  0.0007940  0.0088568 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.6583550  0.0035165 187.221  < 2e-16 ***
log(numcol + 1)      0.7193316  0.0026131 275.284  < 2e-16 ***
log(yieldpercol + 1) 0.0667684  0.0002932 227.732  < 2e-16 ***
log(stocks + 1)      0.0019798  0.0016549   1.196    0.232    
log(prodvalue + 1)   0.0189272  0.0028132   6.728 3.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001766 on 621 degrees of freedom
Multiple R-squared:  0.9996,	Adjusted R-squared:  0.9996 
F-statistic: 3.997e+05 on 4 and 621 DF,  p-value: < 2.2e-16
```

Residual Vs Fitted Plot:

Scatter plot with horizontal lines indicates how the predicted values have been distributed. The more values at the center, the good the model. From the below plot, some values fall away from the center. Those may be potential outliers. However, except for the outliers the values have been predicted well.



Conclusion:

Honey yield production has been decreasing over the years. The effect of pesticides and hive diseases in the year 2006 has not yet mitigated.Since the production is low and consumption is high, the value of honey has been increasing. However, future works on identifying which state yields more and least help producers to invest in those areas to get better yield.

References:

[1] Kaggle Repository

https://www.kaggle.com/jessicali9530/honey-production#honeyraw_2008to2012.csv