Justin Jiang

STA 3000

May 16, 2022

Final Project

Introduction:

This dataset describes the honey production in America. The data is gathered from the U.S department of agriculture. The dataset gathers the amount of honey produced in the U.S for each state between the years 1998 to 2012. The dataset mostly concentrates on the states being able to produce how much honey and how much money is made from selling honey. The main variables used in this data are:

x-variables:

- numcol (quantitative) - number of colonies
- state (categorical) - state of U.S
- priceperlb (quantitative) - price of honey per pound

y-variables:

- totalprod - total amount of honey produced from number of colonies
- prodvalues - total amount of money made from selling honey

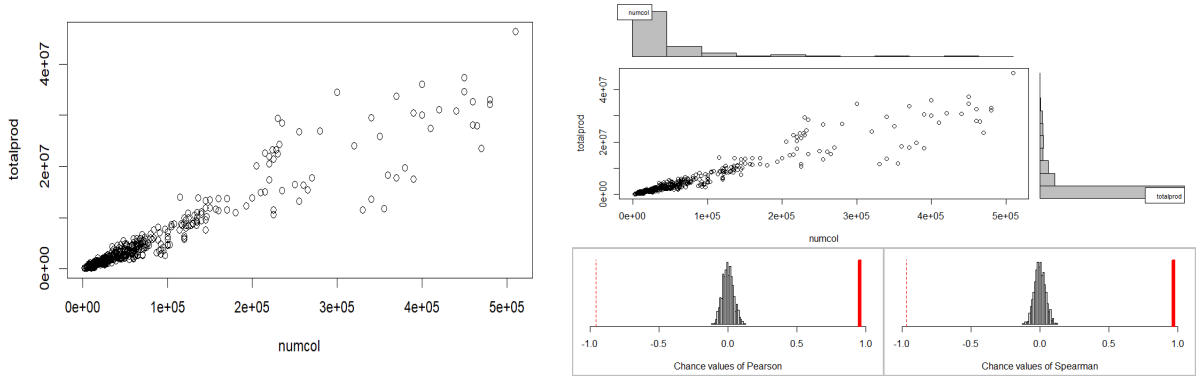
This report aims to see if there is any statistical significance between the location of the states and the amount of honey produced. Also, this dataset explores any statistical significance between the number of colonies and price of honey to the amount of honey produced.

Association Analysis:

For our association analysis the variable that will be predicted is the total amount of honey produced (totalprod). The state, number of colonies (numcol), and the price of honey per pound

(priceperlb) will be the x variables while the total amount of honey produced will be our y variable (totalprod).

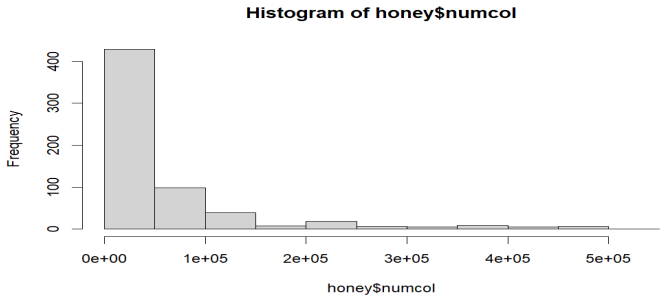
Numcol:

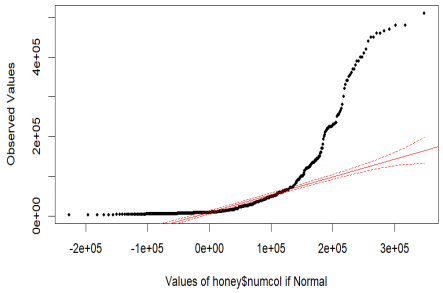



```
Association between numcol (numerical) and  totalprod (numerical)
 using 626 complete cases
Permutation procedure:
                              Value Estimated p-value
Pearson's r                 0.9535944                 0
Spearman's rank correlation 0.9700036                 0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
Note:  If 0.05 is in this range, increase the permutations= argument.
```

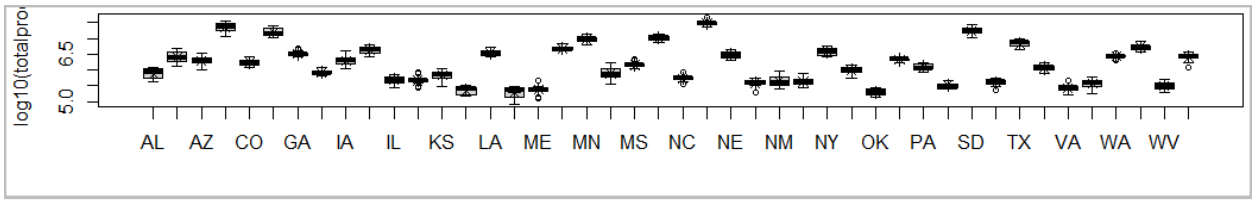
The association between number of colonies and total production of honey seems to appear to have a heteroscedastic relation as the data seems to scatter as the number of colonies increase. Also, according to the qqplot and the histogram using number of colonies as our x variable, a median test should be used because the data in the is outside of the red lines and the distribution of the frequency table is not normal. The r value according to spearman's rank is 0.97 indicating a strong association between the number of colonies and total amount of produce. According to the permutation test, the p-value is 0 making our data statistically significant, making the relation between the number of colonies and the total amount produced to be unlikely to happen by chance. Furthermore, according to spearman rank, the p-value is between 0 and 0.004. A reason

why the number of colonies and total production of honey has a strong association is because the total production of honey is dependent on the number of colonies.



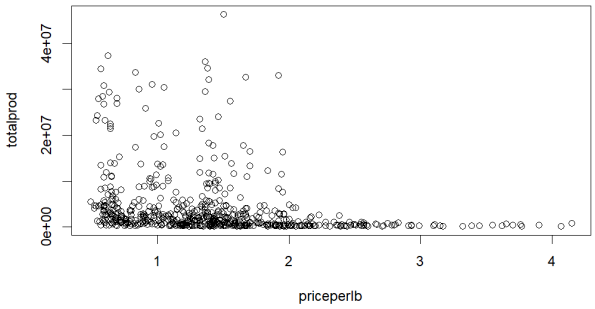


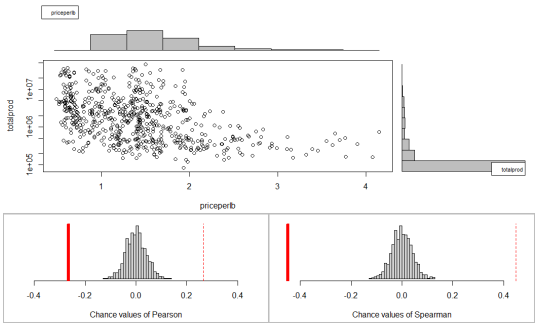
State:



Due to many states, a proper association analysis cannot be used. However, under the associate command, it did give a box plot and so a reasonable analysis can be made. A few states that stand out are California, Florida, North Dakota and South Dakota. There is association with the 4 states because the median and mean are different to each other. The environments of the four states are suitable to raise bee colonies because those states contain a good amount of flora and have warm temperatures. Honey production from other states seem to have similar median and mean.

Priceperib:

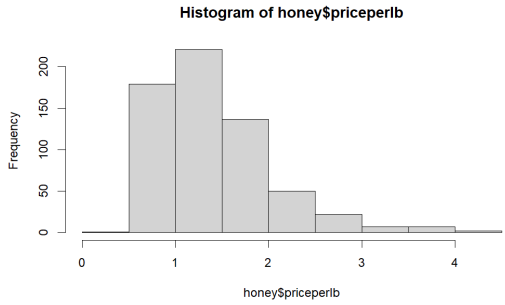


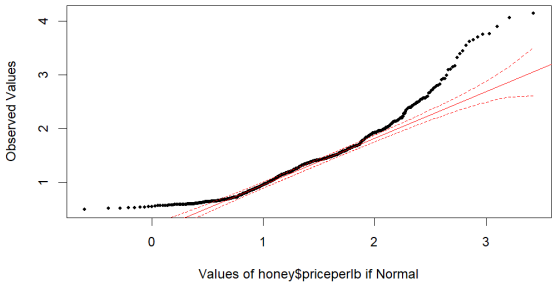



```
Association between priceperlb (numerical) and  totalprod (numerical)
 using 626 complete cases
Permutation procedure:
                              Value Estimated p-value
Pearson's r                -0.2644986                 0
Spearman's rank correlation -0.4491594                0
With 1000 permutations, we are 95% confident that:
 the p-value of Pearson's correlation (r) is between 0 and 0.004
 the p-value of Spearman's rank correlation is between 0 and 0.004
Note:  If 0.05 is in this range, increase the permutations= argument.
```

The association between the price of honey and total production of honey seems to appear to have a weak relation as the line data seems to form a cloud shape as the price decreases. Also, according to the qqplot and the histogram using the price of honey as our x variable, a median test should be used because the data in the is outside of the red lines and the distribution of the frequency table is not normal. The r value according to spearman's rank is -0.44 indicating a weak association between the price of honey and total amount of produce. According to the permutation test, the p-value is 0 making our data statistically significant, making the relation between the number of colonies and the total amount produced to be unlikely to happen by chance. Furthermore, according to spearman rank, the p-value is between 0 and 0.004. A reason why the price of honey and total production of honey has a weak association is because the price of honey might not really entail that there is a lower production of honey. There might be a lurking variable to cause this negative association such as the demand of honey. As shown in the

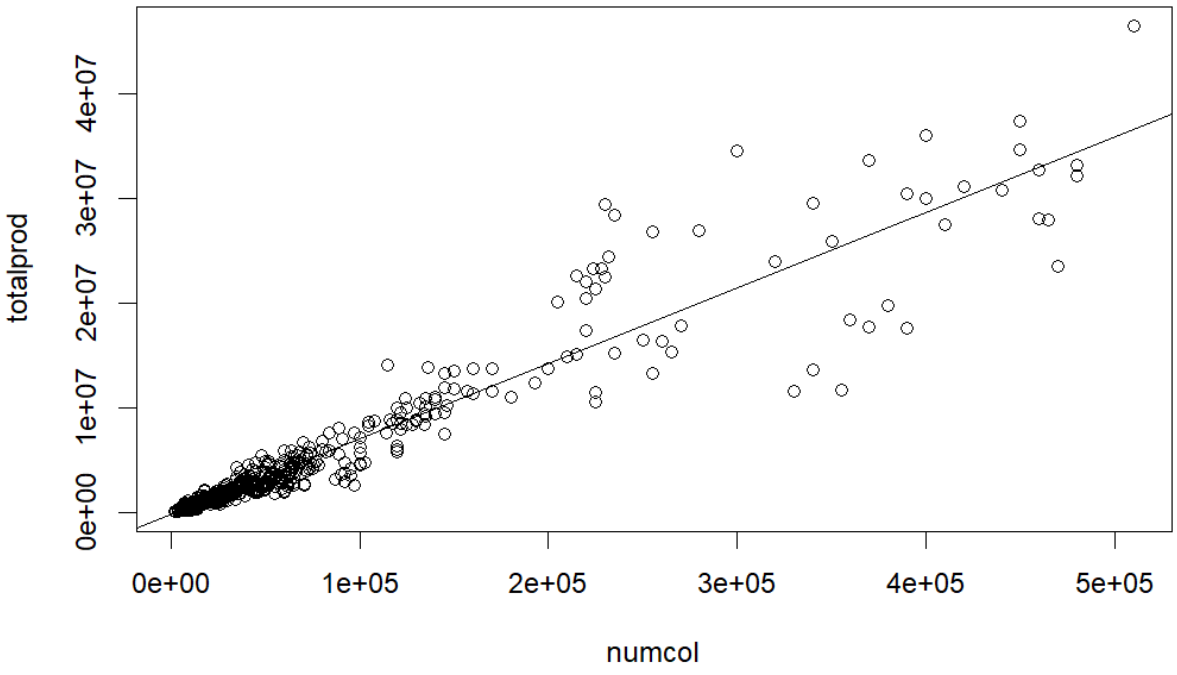
data, lower honey prices do not imply higher honey prices from the data shown; some prices also show lower honey production.

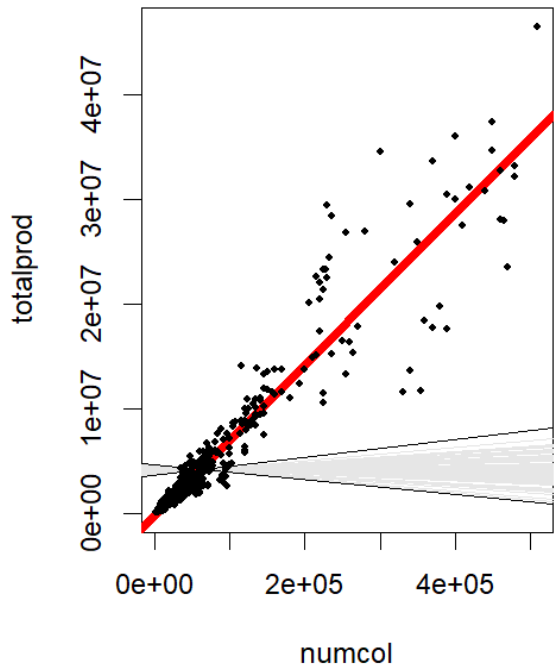


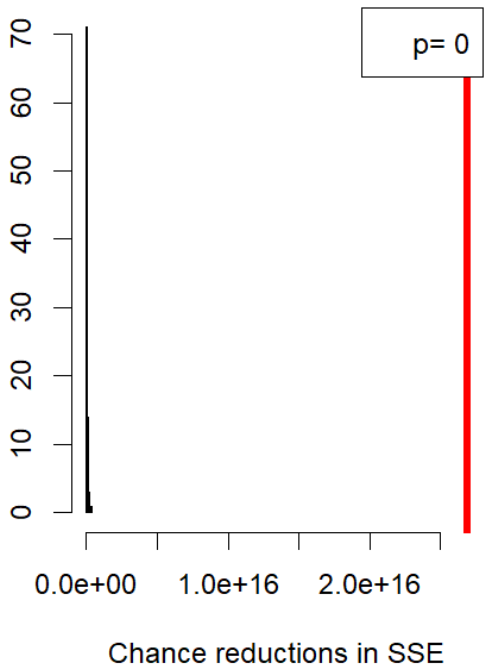


Linear Regression:

When doing a linear regression between the number of colonies and the total amount of honey produced, these are the results shown:







In a relation between the number of colonies and the total amount of honey produced, the best regression equation between the two variables is:

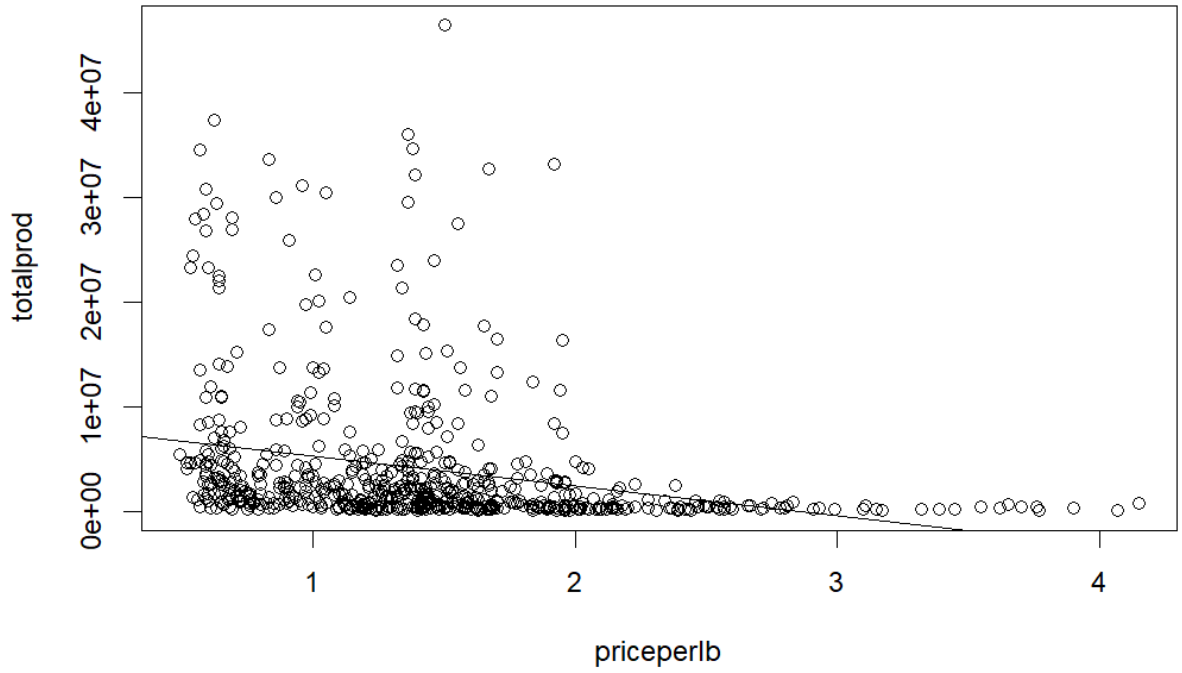
$$y = 72.08x + 175,198.85$$

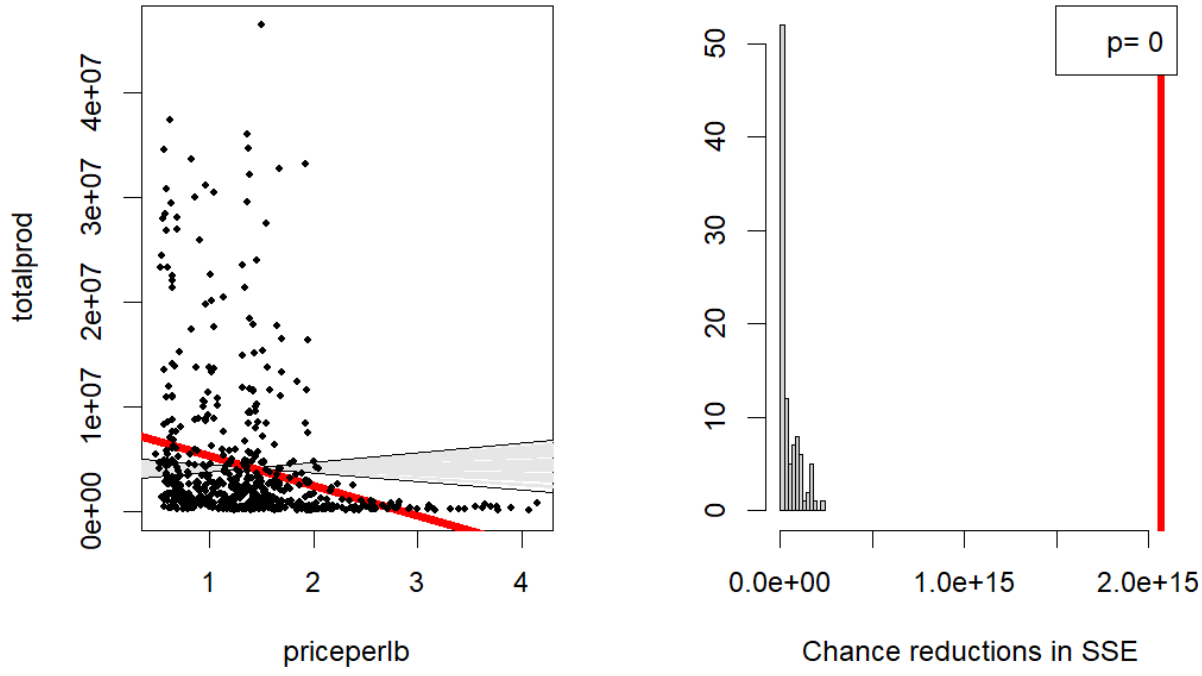
The slope is 72.08 while the y-intercept is 175,198.85. The p- value of this relation is 0 meaning that our data is statistically significant and most likely that the relation between the two is unlikely to happen by chance. This is also shown in the F test where the p-value is 2×10^{-16}, which is less than 5%. The root mean squared error (RSME) is 2,074,000 on 624 degrees of freedom, meaning that there is a 2,074,000 percentage error when predicting the amount of honey produced, meaning that it is not very accurate predicting the total amount produced. The R^2 on the other hand is 90.92%, showing that the number of colonies can predict 90.92% of the amount of honey produced since SSE (sum of squared errors) is reduced by 90.2% from the linear model. Further evidence can be shown where the points are close to the line of the

regression equation. The standard error in the regression model is 0.911. In a 95% confidence interval, the slope is between 70.28 and 73.86. This means that we are 95% confident that the slope is between 70.28 and 73.86. A linear regression model can best describe the relation between the number of colonies and total amount of honey produced since there is statistical significance and the model can predict the amount of honey produced.

```
Call:
lm(formula = totalprod ~ numcol, data = honey)

Residuals:
      Min        1Q    Median        3Q       Max
-13695771   -362972      -183    229154  13053364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.759e+05  9.945e+04  -1.769   0.0774 .
numcol       7.208e+01  9.110e-01  79.114   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2074000 on 624 degrees of freedom
Multiple R-squared:  0.9093,    Adjusted R-squared:  0.9092
F-statistic:  6259 on 1 and 624 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: totalprod
           Df     Sum Sq    Mean Sq F value    Pr(>F)
numcol      1 2.6932e+16 2.6932e+16    6259 < 2.2e-16 ***
Residuals 624 2.6850e+15 4.3029e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                   2.5 %      97.5 %
(Intercept) -371212.68364 19374.98481
numcol           70.28613    73.86424
```

The other quantitative x-variable priceperlb shows these graphs when a linear regression is used to see the relation between the price of honey and the total amount produced:





From the graphs shown, the best line for the linear regression between the two variables is:

$$y = -2851192x + 8188037$$

The slope is -2,851,192 and the y intercept is 8,188,037. The p- value of this relation is 0 meaning that our data is statistically significant and most likely that the relation between the two is unlikely to happen by chance. This is also shown in the F test where the p-value is 1.76×10^{-11}, which is less than 5%. The RSME is 6,644,000 on 624 degrees of freedom, meaning that there is a 6,644,000 percentage error when predicting the amount of honey produced based on the price of honey, meaning that it is not very accurate predicting the total amount produced. The R^2 is 6.99%, showing that the price of honey can predict 6.99% of the amount of honey produced since the SSE is reduced by 6.99% in a linear model. The standard error in the regression model is 416,161. In a 95% confidence interval, the slope is between -3,668,438 and -2,033,945. This means that we are 95% confident that the slope is between -3,668,438 and -2,033,945. A linear regression model is not the best to describe the relation between the price of honey and total amount of honey produced. There is statistical significance, but the model is not the best at predicting the amount of honey produced.

```
Call:
lm(formula = totalprod ~ priceperlb, data = honey)

Residuals:
     Min       1Q   Median       3Q      Max
-6122858 -3427301 -2092303   251978 42498751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8188037     643913  12.716  < 2e-16 ***
priceperlb  -2851192     416161  -6.851 1.76e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6644000 on 624 degrees of freedom
Multiple R-squared:  0.06996,   Adjusted R-squared:  0.06847
F-statistic: 46.94 on 1 and 624 DF,  p-value: 1.76e-11

Analysis of Variance Table

Response: totalprod
             Df     Sum Sq    Mean Sq F value   Pr(>F)
priceperlb    1 2.0720e+15 2.0720e+15  46.938 1.76e-11 ***
Residuals   624 2.7545e+16 4.4143e+13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
                2.5 %   97.5 %
(Intercept)  6923537  9452536
priceperlb  -3668438 -2033945
```

Work Cited

Li, Jessica. *Honey Production in the USA (1998-2012)*. Kaggle, https://www.kaggle.com/datasets/jessicali9530/honey-production