University of Manitoba

Written Report

Submitted By:

Amanpreet Singh (7873300)
Jaspreet Singh (78597906)

Dataset Description

Our team decided to work on the deforestation dataset for the term project. The Dataset contains information about the forest loss worldwide as well as in different countries over the past decades. Furthermore, it provides information on different aspects that may have led to the global forest loss. The dataset was published by "Hannah Ritchie and Max Roser (2021)" at OurWorldInData.org.

Overall, data is separated into five different data-frames each containing information about a specific aspect that might have led to the global forest loss.

The five data-frames are:

1. forest: This dataframe contains information about the change of forest area in different countries every 5 years over the past decades.

2. forest_area: This dataframe contains information about the change in global forest area in different countries as a percentage of global forest area.

3. Brazil_loss: This dataframe contains information about the different causes of forest loss in Brazil every year from 2001 to 2013.

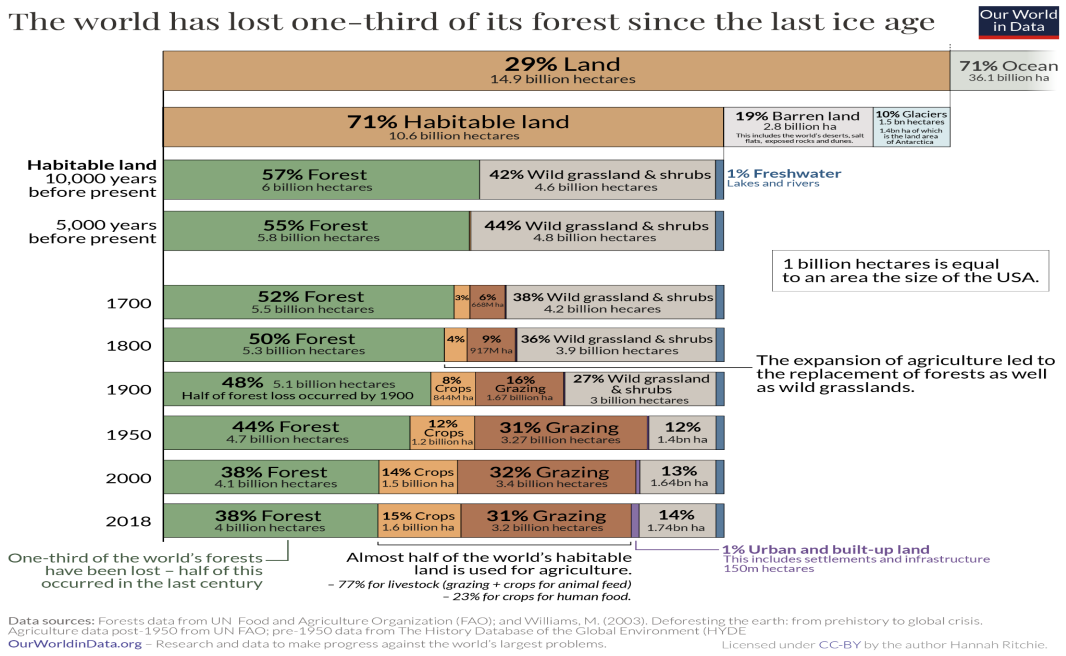
4. soybean_use: This dataframe contains information about soybean production of different countries every year from 1961 to 2013.

5. vegetable_oil: This dataframe contains information about the oil production by the crop type of different countries every year from 1961 to 2013 .

Overview

Our world has experienced a drastic amount of forest loss over the past centuries. Although there are many explanations accounting for humanity's destruction of nature, the population growth and increased consumption concerns the most. According to the estimates, since the 18th century total world population rose from about 600 million to about 7.9 billion as of 2021.

As a result of the global population increase, meeting accommodation, food and fuel needs of humans have received a priority over Environmental protection. There is a false belief that the urban lifestyle and the megacities are one of the causes of deforestation, but in fact urban land only accounts for 1% of global habitable land. So what exactly is the driving factor of global deforestation?



Source: Food and Agriculture Organization of the United Nations (FAO) (2020).

According to the chart published by "Hannah Ritchie and Max Roser (2021)", in 2018 almost 50% of Earth's habitable land was used for agriculture. Thus there is a lot of evidence to prove that agricultural land expansion was the biggest driver of forest loss. In this report we tried to use analytical reasoning to show that agricultural land was in fact the biggest driver of forest loss.

Literature review

There is a plethora of literature relevant to deforestation around the world, plenty of which also relates to our tentative analysis question. For this report our team used the deforestation dataset to perform different statistical analysis, along with two other sources for relevant graphs and literature.

1. Article titled “Forests and deforestation” by Hannah Ritchie and Max Roser (2021): Along with the deforestation dataset which is the primary source of analysis for this report, this article also discusses the change in global agriculture and other causes of forest loss over the past centuries.
2. Article titled “Deforesting the earth from prehistory to global crisis” by Food and Agriculture Organisation of the United Nations (FAO) and Williams M. (2003) discusses the use of Earth’s habitable land by humans and relates it to the causes of forest loss. The article contains specific data visualizations relevant for this report.

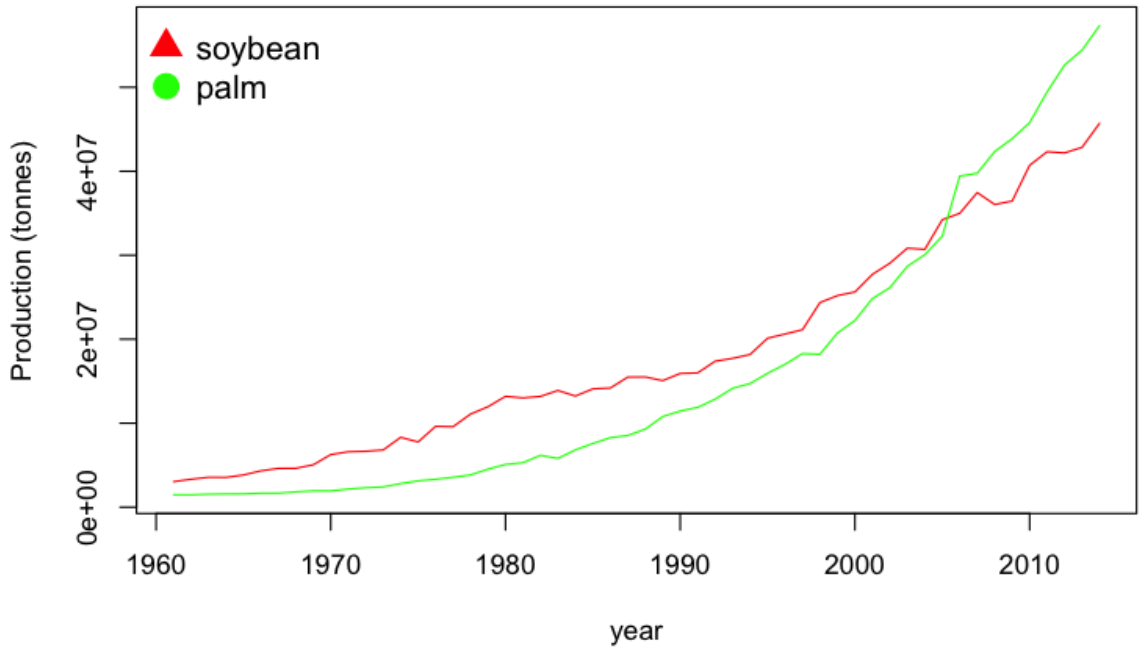
Tentative analysis question

"Was the global forest loss a direct result of Expansion of agricultural land?"

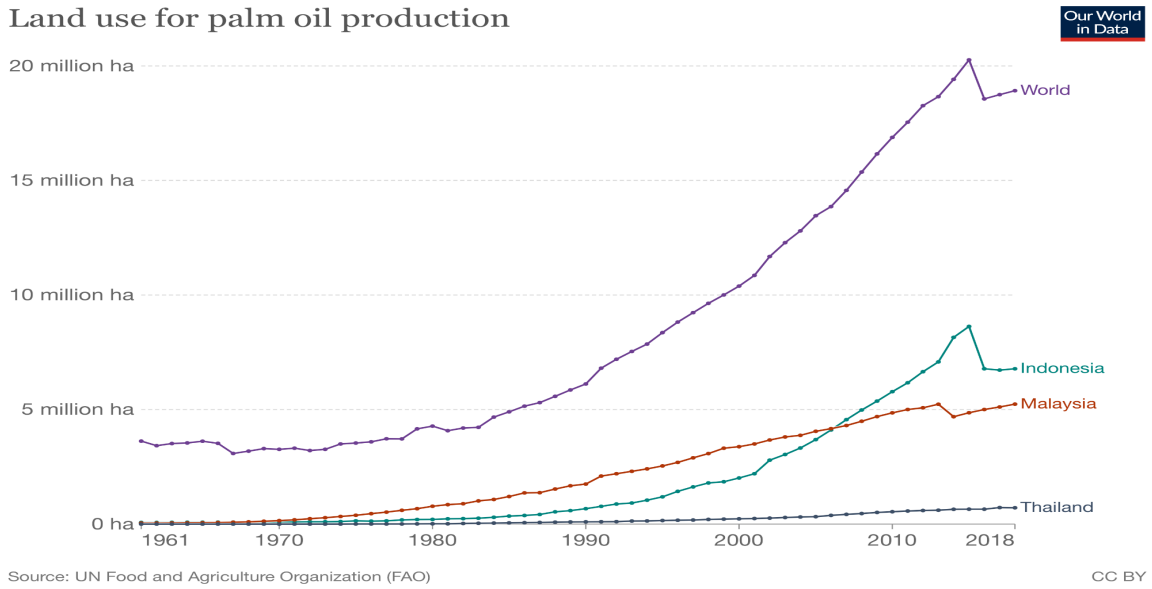
Statistical Analysis

1. Visualizations

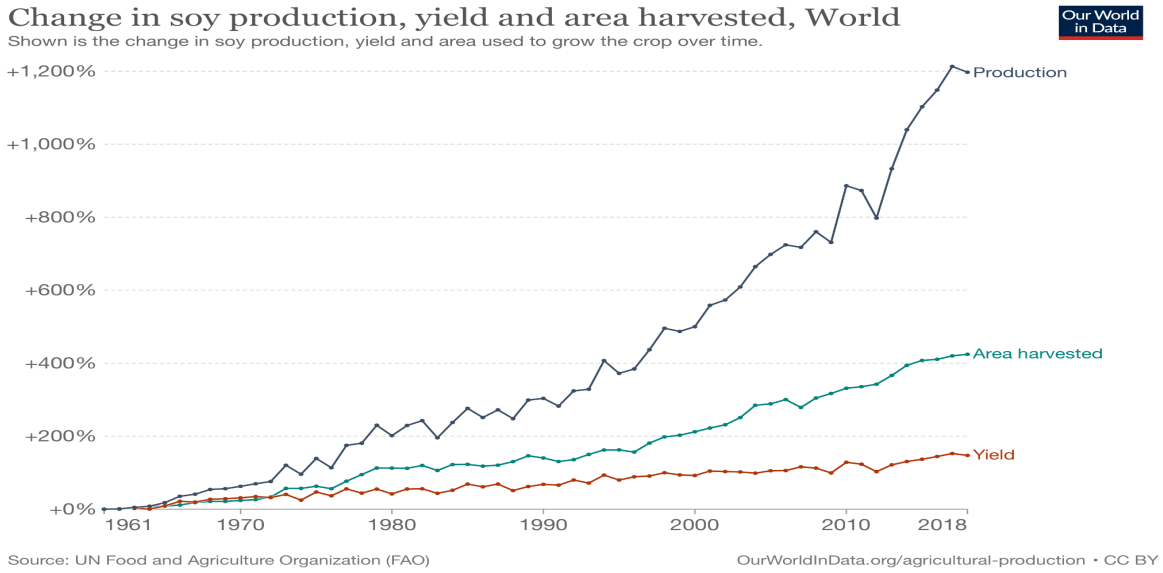
To begin with, the global production of soybean and palm oil experienced some enormous gains as depicted in the chart below. In order to realise the relation between the growth in production and the global forest loss, we must understand how the world increased production in such huge amounts over a short period. In 1960, the world produced around 20 to 30 million tonnes each of soybean and palm which increased more than ten fold by 2018. Either higher yields or agricultural land expansion can justify this increase.



Although recent developments in farming techniques have resulted in increased crop yields, cropland expansion was necessary to keep up with the increased production. The two charts below were published by the, displaying the agricultural land expansion in relation to the increased production of soybean and palm oil. Since 1960 land use for palm oil production has quadrupled and the land used for soy production also increased significantly. Most of the expansion in land used to produce soybean and palm came at an expense of forest area, thus proving that the agricultural land was a direct driver of forest loss.



Source: Food and Agriculture Organization of the United Nations (FAO) (2020).



Source: Food and Agriculture Organization of the United Nations (FAO) (2020).

2. Correlation

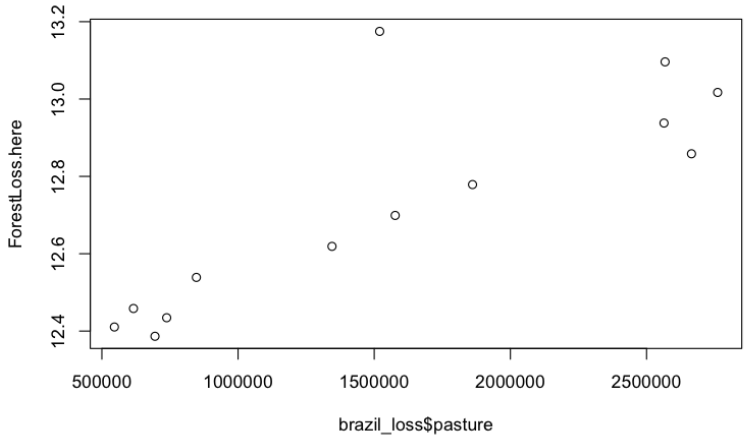
Another way to see how agricultural land expansion is the leading cause of forest loss is to check the correlation between forest area lost and the different causes. In statistics, correlation defines the relationship between different variables i.e how a change in one variable might affect the other.

Here we will define a linear relationship on different causes of forest loss in Brazil and the total forest area loss in Brazil over the years using Pearson correlation. Pearson correlation provides us with values $\in$ [-1, 1], +1 being a strong positive relation and -1 being a strong negative relation.

```
commercial_crops                0.72932842
flooding_due_to_dams            0.37140188
natural_disturbances           -0.46392258
pasture                         0.83463380
selective_logging               0.06697744
fire                           -0.16125398
mining                         -0.20671629
other_infrastructure            0.34138127
roads                           0.29229231
tree_plantations_including_palm 0.48723141
```

Although infrastructure and roads had a significant correlation with forest area loss, it is evident that the commercial crops and the pasture were most correlated to the forest area loss in Brazil. Both commercial crops and pasture fall in the category of agricultural land use, thus it is safe to say that the result is in support of our claim.

To better establish this correlation it is also useful to look at the relationship on a plot.



The scatter plot clearly shows a somewhat linear relation between the forest loss in Brazil with respect to the increase in land used for pasture which again falls under the category of agricultural land use.

3. Hypothesis Testing

Another technique to study the relation between agricultural land expansion and forest loss around the world is hypothesis testing. In statistics, hypothesis testing is an important method to prove or disprove an assumption.

Although hypothesis testing is a great method to study relations, the answer is based on probability and might not be certain which is certainly a limitation.

Let's assume that our null hypothesis is that "there exist no relation between forest loss and commercial crops yield" using Brazil_loss dataset. With $\alpha = 0.05$ level of significance we will try to reject the null hypothesis, hence proving that the forest loss is related to the commercial crop yield.

The resulting P-value after running a t_test on the data is 0.002353 which is smaller than $\alpha = 0.05$ level of significance, thus we can reject the null hypothesis, so there exists a relation between forest loss and commercial crops yield.

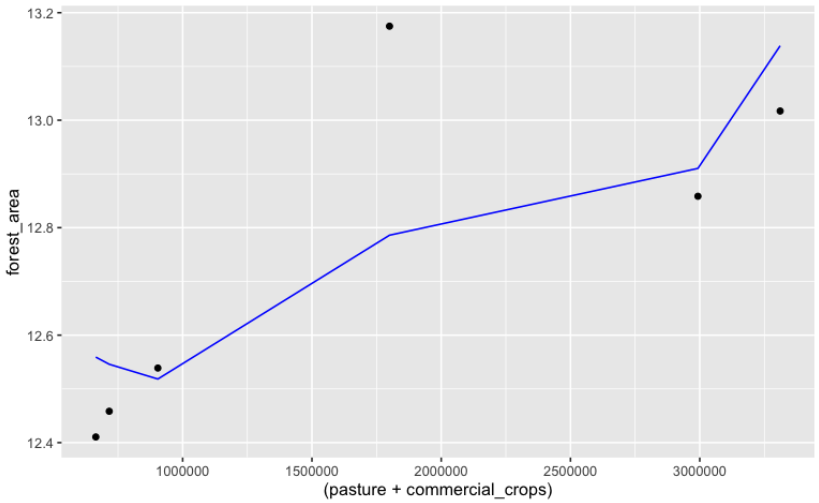
Computing the same calculation on land used for “pasture”, we found the p-value is 2.459e-05 which again rejects the null hypothesis and we can see that the land used for pasture is also related to forest loss.

Both pasture and commercial crop fall in the category of agricultural land use. Having established a relationship between these two and the forest loss, by the use of hypothesis testing we proved that the forest loss is related to the agricultural land expansion.

4. Linear regression

In statistics linear regression is a way to study correlation and to form prediction models. A simple linear regression model is a mathematical equation that allows us to predict a response for a given predictor value. For this report, we will split the given data into training and test data and check how close we can predict to the actual value by building models using linear regression.

Suppose we want to predict the forest loss in Brazil using the land used for commercial crops and pasture . This technique will help us establish a correlation between the forest loss and agricultural land expansion.



Due to the lack of data our model had a limitation but it was still able to perform some close prediction. The rmse of our linear model is 0.172 which seems to be quite reasonable given that the training data was limited and the data being taken over a large period, there can be many outliers which impacts the performance of our model.

5. Decision Tree Regression

In Statistics, Decision trees are a lot common in studying relationships between variables. Decision trees are also a vital part of machine learning algorithms. In this report, our team used a decision tree algorithm to apply regression on the dataset and we tried to predict the forest area loss using the data depicting the land used for agriculture.

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

In [20]:
```
X = [1520000,2568000, 2761000, 2564000, 2665000, 1861000, 1577000, 1345000,  847000,  616000,  738000,  546000, 695000]
```

In [21]:
```
y = [13.17491, 13.09609, 13.01708, 12.93786, 12.85845, 12.77884, 12.69903, 12.61901, 12.53880 ,12.45838, 12.43449, 12.41055, 12.38656]
```

In [48]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.05)
```

In [49]:
```
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor()
regressor.fit(np.array(X_train).reshape(-1,1) , np.array(y_train).reshape(-1,1) )

y_pred = regressor.predict(np.array(X_test).reshape(-1,1))
y_pred
```

Out[49]:
```
array([12.93786])
```

In [50]:
```
df = pd.DataFrame({'Real Values':np.array(y_test).reshape(-1), 'Predicted Values':np.array(y_pred).reshape(-1)})
df
```

Out[50]:

	Real Values	Predicted Values
0	13.09609	12.93786

In the given code variable X is the land used for agriculture and variable y is the forest area lost. Furthermore, the final prediction of the Decision tree model is very close to the real value which depicts a strong relation although it is important to account that a limited amount of data was available to perform calculations which certainly have an impact on the accuracy of the model.

Discussion/Conclusion

The past few centuries brought some major changes in human life. The ever growing notion of globalization has significantly enhanced the exchange of resources between different countries around the world. Although globalization has many benefits, the

seemingly exponential population growth over the past centuries has increased the demand for many resources like food, clothing and shelter. Furthermore, this increased demand pushes for an increase in production of these resources. The countries and the big firms have been trying so hard to keep up with demand that they hardly thought of the sustainable way to move forward.

In this report our team performed an analysis of unsustainable use of Earth's habitable land and forest area. In the overview of this report, we stated that about 50% of Earth's habitable land is used for agricultural purposes and just about 2 centuries ago it was only 10%. Most of this increase in agricultural land came at an expense of forest area.

We used many different statistical approaches to perform analysis on the given dataset and the results were all in support of the fact that the unsustainable expansion of agricultural land is driving forest loss around the world. The methods used in this report include data visualization, correlation, hypothesis testing and linear regression.

At last, in accordance with the outcome of the methods used in this report our team concludes that "expansion of agricultural land is driving forest loss around the world".

Limitations

The data used in some of the methods come from different datasets written over different time periods which makes it hard to organize. This limits the performance of these methods and can provide biased results.

References

Hannah Ritchie and Max Roser (2021) - "Forests and Deforestation". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org'

Williams M. (2003) "Deforesting the earth from prehistory to global crisis". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org'

Appendix

R Code:

```
forest <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-06/forest.csv')
forest_area <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-06/forest_area.csv')
brazil_loss <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-06/brazil_loss.csv')
soybean_use <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-06/soybean_use.csv')
vegetable_oil <-
readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-06/vegetable_oil.csv')

library(tidyverse)
dataset = pivot_wider(forest, id_cols = "year", names_from = "entity", values_from = "net_forest_conversion")
world.data = -dataset[which(colnames(dataset)=="World")]
plot(dataset$year, t(world.data), type = 'l',xlab = "year", ylab = "forest area lost")

data1 = forest[forest$year=="2015",]
data2 = forest_area[forest_area$year =="2015",-c(2,3)]

data = inner_join(data1, data2, by = "entity")
data = data[order(data$forest_area, decreasing = TRUE),]

sum(data[2:6, 5])
sum(data[2:6, 4])/as.numeric(data[1,4])*100
cor(data$forest_area, -data$net_forest_conversion)

data_brazil = forest_area[which(forest_area$entity %in% c("Brazil")),]
ForestLoss.here = data_brazil[which(data_brazil$year %in% brazil_loss$year),]$forest_area
cor(brazil_loss[ , 4],ForestLoss.here)
plot(brazil_loss$pasture,ForestLoss.here)

data.soy = soybean_use %>% drop_na()
data.soy$total = data.soy$processed +data.soy$animal_feed +data.soy$human_food
data.soy = pivot_wider(data.soy, names_from = "year", id_cols = "entity", values_from = "total")[49,]
```

```
plot(colnames(data.soy[2:54]), data.soy[2:54], type = 'l',xlab = "year", ylab = "soy
production")

data.oil = vegetable_oil[which(vegetable_oil$entity == "World"),]
data.oil = data.oil[which(data.oil$year %in%
c("2005","2006","2007","2008","2009","2010","2011","2012","2013","2014")),]
data.oil = pivot_wider(data.oil, id_cols = "crop_oil", names_from = "year", values_from =
"production")

plot(colnames(data.oil[,2:11]),data.oil[1,2:11], type = "l", col = "red",xlab = "year", ylab =
"Production (tonnes)", ylim = c(10723821,57328872))
lines(colnames(data.oil[,2:11]),data.oil[4,2:11] , col = "green")
lines(colnames(data.oil[,2:11]),data.oil[5,2:11] , col = "blue")
lines(colnames(data.oil[,2:11]),data.oil[11,2:11] , col = "yellow")
legend("topleft",
  legend = c("soybean", "palm","Rapseed", "sunflower"), col = c("red", "green","blue",
"yellow"),
  pch = c(17,19),
  bty = "n",
  pt.cex = 2,
  cex = 1.2,
  horiz = F )

newData = data.frame(pasture = as.data.frame(brazil_loss[ ,
1:13])$pasture,commercial_crops = as.data.frame(brazil_loss[ , 4:13])$commercial_crops,
forest_area = data_brazil[12:24,]$forest_area)
train_test =
c("train","test","train","test","train","test","test","test","train","train","test","train","test")
newData = mutate(newData, train_test)
data_test = newData[which(newData$train_test=="test"),]
data_train = newData[which(newData$train_test=="train"),]
fit1 = lm(forest_area~pasture+ commercial_crops, data = data_train)
predict1 = predict(fit1, newdata = data_test)
rmse1 = sqrt(mean((data_test$forest_area-predict1)^2))
data_train %>% mutate(.fitted = fitted(fit1)) %>% ggplot(aes(x =
(pasture+commercial_crops)))+geom_point(aes(y=forest_area))+geom_line(aes(y=.fitted),co
lour = "blue")
```