Data Analysis Report - Holy Crop!

Hailey Shaver, Jacob Baum, Gaurav Law, and Ben Strong

2024-04-24

Abstract

This report explores the effects of land and fertilizer on total crop yield across many different regions. There were many variables considered in the analysis of crop yield. We chose to look at the Entity, amount of fertilizer used, and the amount of arable land available to grow crops and to observe how each of these affected individual crops as well as the total yield of all the cereal crops. Some of the main regions we analyze the effects of are the Americas (North and South), Asia (Southern, Western, and South Eastern), Australia and New Zealand, the Caribbean, Central America, Africa (Eastern, Western, and Southern), Europe (Eastern, Western, and Southern), Oceania, To study these effects, we utilize logistic regression analysis to quantify the effects of these variables on the crop yield. By using our logistic regression model, we are able to compare the changes in yield based on the different factors for any given year as well as predict a potential cereal crop yield for future years based on our data. Through these findings, we can quantify the importance of fertilization management and strategies to increase productivity and yield of crops across many different countries and regions.

After testing many different logistic models, we came to find out that arable land provided the least effect on the crop yield which was very interesting as it was speculated to have been one of the bigger contributing factors to total yield. Fertilizer had the highest positive correlation with respect to the crop yield and had a more dramatic effect on the total amount of crops produced. The crops that had the most significant increase in yield (with respect to the p-value) due to the amount of fertilizer used were rice, potatoes, cassava, and beans.

Introduction

Crop yield is our main area of concern, and the problem that arises out of this is what conditions and factors can affect it given data recorded from previous years. What regions around the world are affected differently than others, naturally was one of the first areas that we thought to look at. Which crops have the largest impact on the yield overall, also

appeared as we were analyzing the data. How has nitrogen fertilizer affected crop outputs and changed over the years.

To answer these questions, we created multiple logistic regression models to study the strength of each variable's effects on both individual crops and the total crop yield. With these models, we compared the AIC and p-values of each covariate and its effects on the yield. After careful analysis and visualizations of the different variables and their effects, we noticed that our original assumptions about region/Entity having a relatively great effect on crop yield was not as significant as we had first thought. However, fertilizer did in fact have a larger impact on each individual crop yield, as well as the overall cereal yield which confirmed our second assumption. We chose to look at many different factors and covariates for why our original assumption about the region having an impact on crop yield/ overall cereal yield wasn't confirmed and what reasons there could be behind that. We investigated many different models to attempt to see the effects, and we found that the quasi-poisson model most accurately depicts the interactions of multiple variables on the total crop yield.

Data and Methods

The data that we used was the Global Crop Yields from Our World In Data, through the tidytuesday github. There were five datasets in total and all share the categorical variables of Entity (Country/Region Name), Code (Country Code) and Year. The first dataset was focused on yields of key crops in tonnes per hectare such as wheat, rice, maize, etc. The second dataset included the arable land needed to produce a fixed quantity of crops normalized to 1961. The third dataset included nitrogen fertilizer usage in kilograms per hectare and cereal yields in tonnes per hectare. The fourth dataset focused on cereal yield index with a variable for the change to land area for cereal production since 1961, as well as population. The fifth data set looked at tractors per 100 sq km of arable land, as well as cereal yield in kg per hectare and the total population as well.

We decided to only focus on the first three datasets to cut down on the number of variables, then to clean the data further we filtered out the data to only include larger Entities, reducing our data to large regions/continents so we could use Entity as a categorical variable with a much smaller number of categories. Then after merging the filtered crop yields, arable land and fertilizer datasets we started looking at the year variable. Some Entities did not have values for certain years, so we had to limit our models to only use a certain range of years, where we settled on 1995-2014. There were still some Entities that did not have one specific type of crop, but these were ignored when we made the models.

After cleaning the data, we classify our outcomes as the total yield of crops in kilograms per hectare and denote the observed values as yi where i = 1, ..., n, n = 460. We then started searching for models using stepwise regression starting with the full model, removing some variables to minimize the AIC. At this point we split our focus to creating two models, one that focuses on the effects of arable land, Entity and specific crops over time on total yield, and another that looks at the effect of fertilizer on total yield.We also found that a quasi-poisson model was able to show us better models when looking at the residuals

For the Entity model we assumed that each logit(p_i) is a realization of a random variables $Y_i \sim$ Quasi-Poisson(λ_i) independently, and model the count of total yield using the following logistic regression with these variables:

$$\begin{aligned}\text{logit}(p_i) =& \beta_0 + \\ & \beta_1 \times (\text{Entity})_i \\ & \beta_2 \times (\text{Barley})_i \\ & \beta_3 \times (\text{Cocoa})_i\end{aligned}$$

Where $i = 1, ..., n$ and $n = 440$.

For the Year model we assumed that each logit(p_i) is a realization of a random variables $Y_i \sim$ Quasi-Poisson(λ_i) independently, and model the count of total yield using the following logistic regression with these variables:

$$\begin{aligned}\text{logit}(p_i) =& \beta_0 + \\ & \beta_1 \times \text{I}(\text{Entity})_i \\ & \beta_2 \times (\text{Wheat})_i \\ & \beta_3 \times (\text{Maize})_i \\ & \beta_4 \times (\text{Soybeans})_i \\ & \beta_5 \times (\text{Potatoes})_i \\ & \beta_6 \times (\text{Cassava})_i \\ & \beta_7 \times (\text{Barley})_i \\ & \beta_7 \times (\text{Cocoa})_i\end{aligned}$$

Where $i = 1, ..., n$, $n = 440$ and I() in the equation indicates a dummy variable, taking value 1 when the condition is true and zero otherwise.

For the Arable Land model we assumed that each logit(p_i) is a realization of a random variables $Y_i \sim$ Quasi-Poisson(λ_i) independently, and model the count of total yield using the following logistic regression with these variables:

$$\begin{aligned}\text{logit}(p_i) =& \beta_0 + \\ & \beta_1 \times (\text{Land})_i \\ & \beta_2 \times (\text{Wheat})_i \\ & \beta_3 \times (\text{Maize})_i \\ & \beta_4 \times (\text{Potatoes})_i \\ & \beta_5 \times (\text{Cassava})_i \\ & \beta_6 \times (\text{Barley})_i \\ & \beta_7 \times (\text{Cocoa})_i\end{aligned}$$

Where $i = 1, ..., n$ and $n = 440$.

For the Crop model we assumed that each logit(p_i) is a realization of a random variables Y_i ~ Quasi-Poisson(λ_i) independently, and model the count of total yield using the following logistic regression with these variables:

$$\begin{aligned}\text{logit}(p_i) = &\beta_0 + \\ &\beta_1 \times (\text{Wheat})_i \\ &\beta_2 \times (\text{Maize})_i \\ &\beta_3 \times (\text{Soybeans})_i \\ &\beta_4 \times (\text{Potatoes})_i \\ &\beta_5 \times (\text{Cassava})_i \\ &\beta_6 \times (\text{Barley})_i \\ &\beta_7 \times (\text{Cocoa})_i\end{aligned}$$

Where $i = 1, ..., n$ and $n = 440$.

For the Fertilizer model we assumed that each logit(p_i) is a realization of a random variables $Y_i \sim$ Quasi-Poisson(λ_i) independently, and model the count of total yield using the following logistic regression with these variables:

$$\begin{aligned}\text{logit}(p_i) = &\beta_0 + \\ &\beta_1 \times (\text{Fertilizer})_i \\ &\beta_2 \times (\text{Wheat})_i \\ &\beta_3 \times (\text{Rice})_i \\ &\beta_4 \times (\text{Maize})_i \\ &\beta_5 \times (\text{Soybeans})_i \\ &\beta_6 \times (\text{Potatoes})_i \\ &\beta_7 \times (\text{Beans})_i \\ &\beta_8 \times (\text{Cassava})_i \\ &\beta_9 \times (\text{Barley})_i \\ &\beta_1 0 \times (\text{Cocoa})_i \\ &\beta_1 1 \times (\text{Bananas})_i\end{aligned}$$

Where $i = 1, ..., n$ and $n = 440$.

We also took a look at correlation between each of our variables and the yield. The resulting graphs and values showed an extreme amount of correlation between them all. There were a few instances of insignificant correlation, primarily cocoa beans when compared to all other variables, except against arable land, wheat, and bananas.

Results

#Individual Modeling

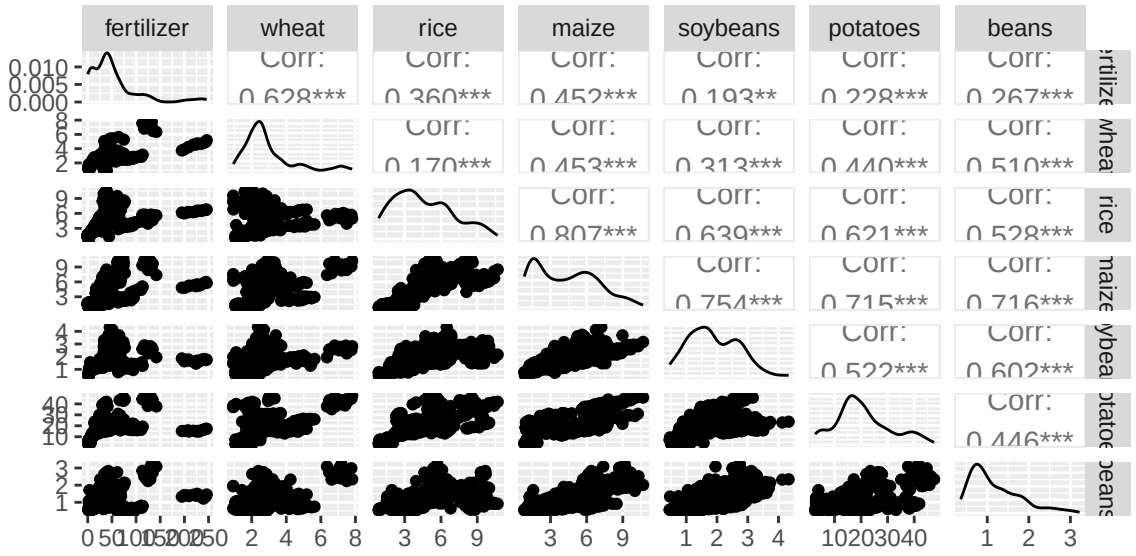

Figure 1: This graph shows and example of the correlation plot values, this particular one has the variables Fertilizer, Wheat, Rice, Maize, Soybeans, Potatoes, and Beans. Basically, all other cordinate plots look similiar to this one.

Overall, we modeled yield as our response variable. We first started with individual variables to determine which might have the largest effect on crop yield. (Findings were excluded if they didn't provide significant effects, i.e. $p < 0.05$.)

Year

Modeling year as a factor variable, we found that a majority of years had significant effects on yield in the quasi-poisson model (negated 1998, 2000, 2005, 2006, 2013, 2014). We did find that the biggest increase in yield came in 2007 on a logistic scale — a change from 1995 to 2007 corresponds to a 1.108-multiplicative odds increase in crop yield. 2001 and 2003 follow closely behind, with 1.099 and 1.043-multiplicative odds increases in yield, respectively. There were no significant multiplicative-odds decreases. Entity

Entity

We decided to not have a baseline for the model consisting solely of entity as the covariate — Africa would have been the baseline, but any other continent with respect to this simply didn't make sense, and we were not getting any significant effects.

Removing Africa as a baseline yielded significant effects for every region studied — this was one of just two models overall which we chose to stick with poisson models instead of quasi-poisson because the dispersion parameter was close enough to 1 that it didn't really make a difference on estimates nor p-values.

Micronesia yielded a 1.147-multiplicative-odds increase in crop yield (remember, there's no baseline so it's not exactly with respect to any other entity). How we interpreted this was it was simply increasing from zero.

Arable land

With arable land, we also removed the baseline because it is difficult to determine what the baseline would even be — maybe non-arable land? When we removed this, we found that there was a significant effect and a 1.409-multiplicative-odds increase in crop yield.

Crops

With each crop, we originally had a baseline, but, similarly to arable land, it was difficult to determine what that was. In addition, all the crops were giving us insignificant effects, leading us to determine removing the baseline was the best way to go. Again, these are just individual crops' effects on yield, these are by no means our best models.

We found that cocoa beans had the highest multiplicative-odds increase of 2.18. This makes sense because, as you will see in other models, this was part of our best crops model. Bananas, on the other hand, had the least positive multiplicative-odds increase, with just 0.036, on yield, which also makes sense because this fruit rarely appeared in any of our good models.

Interaction Effects

Entity*Year

With Africa as a baseline, there was not a single significant effect. Removing the baseline of Africa did not change much for the interaction effects. There were still no significant effects among interactions — all the significant effects with the regular variables were described above.

Entity*wheat

With respect to wheat produced in Africa as a baseline, we found that no positive odds increases with any interactions between wheat and another region. We did find that the biggest change was Western Africa — a change from Africa to Western Africa produces a 0.709-multiplicative odds decrease in yield when only considering wheat as a crop. The region with the best log odds ratio, however, was Southern Africa, which, with respect to Africa, had a 0.269-multiplicative-odds decrease while producing wheat in overall crop yield.

Entity*rice

With respect to rice produced in Africa as a baseline, we once again found that no positive odds increases with any interactions between rice and another region. The most significant log odds ratio change was a one-unit change in rice produced in the Caribbean corresponded to a 0.9745-multiplicative-odds decrease in yield. Perhaps it's not really fair to compare anywhere to Africa in terms of rice production because it has been cultivated for several thousand years, but the Caribbean still produces a good amount of rice today.

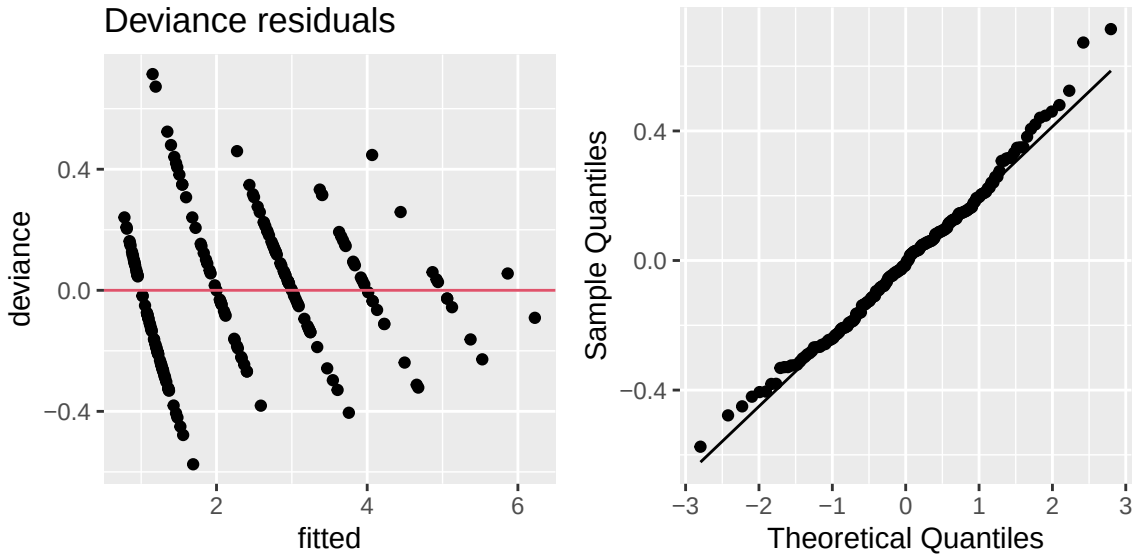

Figure 2: The left graph shows the deviance residuals for Year vs. fitted values and the right shows the Q-Q Normal Plot of the deviance of Year.

The residual plot shows a few things about our model. The first thing to notice is that the values seem to have a constant dispersion meaning that we do have homoscedasticity. There are a few areas where the distribution is a little off of constant, but as a whole. There are, however, a few problems that need to be addressed with the residual plots. The first being a clear pattern in the residual plot, this shows that the relationship between variables and predictors are not linear, it can also mean that the data is not well modeled. Since this exact same pattern repeats over every graph, it seems that the relationship between variables and predictors is more likely than an unfit model. The final problem again occurs because of the bands seen in the plots, it also means that we have not effectively found the underlying cause of the relationship. While these do not bode well for our models, it is imperative for us to state that these graphs are an improvement from previous models we used. We used the Poisson distribution at first and found all the same problems as previously mentioned, but also the dispersion was not constant as well. In looking at these problems we changed to the Quasi-Poisson distribution which fixed our dispersion issue. The QQ-Plots for each model all can be safely assumed to have normal distributions. There are a couple of spots in the plot that trail off the main line a little bit, but generally they stay on the line.

This is the best fitting model using Entity as one of the variables.

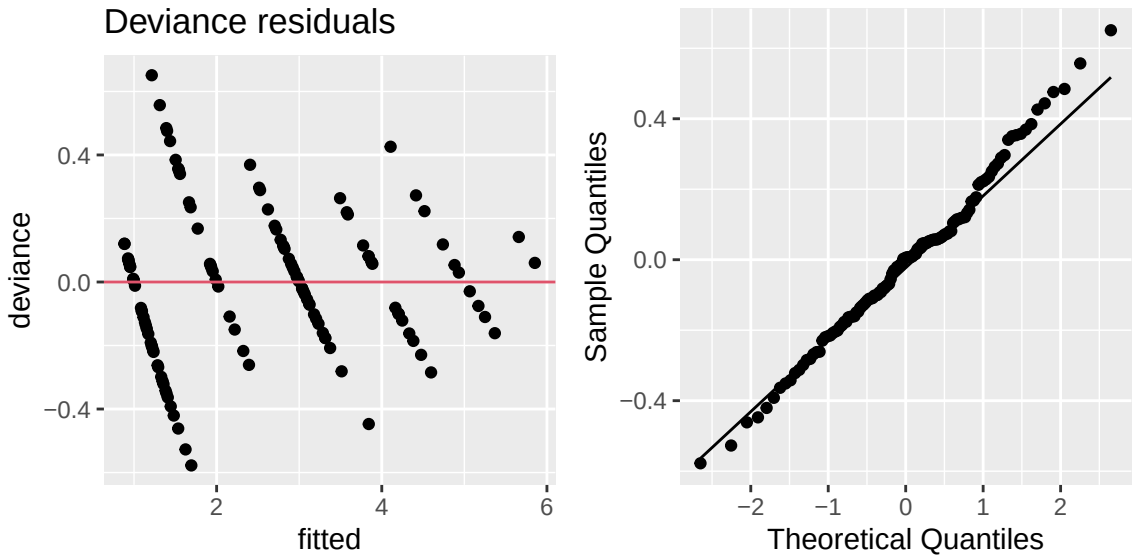

Figure 3: The left graph shows the deviance residuals for Fertilizer vs. fitted values and the right shows the Q-Q Normal Plot of the deviance of Fertilizer.

Standard errors: MLE

	Est.	S.E.	t val.	p
(Intercept)	0.08	0.08	1.03	0.30
EntityAmericas	0.64	0.08	8.00	0.00
EntityAsia	0.88	0.06	14.89	0.00
EntityCentral America	0.45	0.07	6.62	0.00
EntityEastern Africa	-0.07	0.06	-1.08	0.28
EntityMiddle Africa	-0.13	0.07	-1.90	0.06
EntityOceania	0.09	0.06	1.39	0.17
EntitySouth America	0.53	0.07	7.63	0.00
EntitySouthern Asia	0.45	0.06	7.61	0.00
EntityWestern Africa	-0.52	0.08	-6.38	0.00
barley	0.33	0.03	10.57	0.00
cocoa	-0.47	0.12	-3.82	0.00

Estimated dispersion parameter = 0.05

This model uses Year in the model with the best fitting variables to go with it.

Standard errors: MLE

	Est.	S.E.	t val.	p

(Intercept)	-0.94	0.09	-10.13	0.00
as.factor(Year)1996	-0.09	0.08	-1.15	0.25
as.factor(Year)1997	-0.07	0.08	-0.88	0.38
as.factor(Year)1998	-0.18	0.08	-2.33	0.02
as.factor(Year)1999	-0.13	0.08	-1.73	0.09
as.factor(Year)2000	-0.09	0.08	-1.15	0.25
as.factor(Year)2001	-0.13	0.08	-1.66	0.10
as.factor(Year)2002	-0.11	0.08	-1.49	0.14
as.factor(Year)2003	-0.21	0.08	-2.76	0.01
as.factor(Year)2004	-0.14	0.08	-1.79	0.08
as.factor(Year)2005	-0.11	0.07	-1.49	0.14
as.factor(Year)2006	-0.15	0.08	-2.04	0.04
as.factor(Year)2007	-0.18	0.07	-2.36	0.02
as.factor(Year)2008	-0.20	0.07	-2.69	0.01
as.factor(Year)2009	-0.10	0.07	-1.35	0.18
as.factor(Year)2010	-0.15	0.07	-2.02	0.05
as.factor(Year)2011	-0.19	0.07	-2.59	0.01
as.factor(Year)2012	-0.13	0.07	-1.84	0.07
as.factor(Year)2013	-0.08	0.07	-1.16	0.25
as.factor(Year)2014	-0.17	0.07	-2.34	0.02
wheat	0.28	0.02	14.75	0.00
maize	0.18	0.01	12.18	0.00
soybeans	0.32	0.04	8.26	0.00
potatoes	-0.03	0.00	-11.13	0.00
cassava	0.03	0.00	16.98	0.00
barley	-0.02	0.03	-0.62	0.54
cocoa	0.27	0.11	2.53	0.01

Estimated dispersion parameter = 0.06

This is the best fitting model using Arable Land as one of the variables.

Standard errors: MLE

	Est.	S.E.	t val.	p
(Intercept)	-1.92	0.16	-12.08	0.00
land	1.58	0.19	8.36	0.00
wheat	0.36	0.02	17.99	0.00
maize	0.26	0.01	19.57	0.00
potatoes	-0.03	0.00	-11.34	0.00
cassava	0.04	0.00	16.41	0.00

```
barley                0.06   0.03     2.05   0.04
cocoa                 0.34   0.11     3.17   0.00
-----------------------------------------------

Estimated dispersion parameter = 0.06
```

This model uses Crops as the variables without any others, here is what fitetd the data the best.

```
Standard errors: MLE
-----------------------------------------------
                     Est.   S.E.   t val.      p
----------------- ------- ------ -------- ------
(Intercept)         -1.02   0.08   -13.52   0.00
wheat                0.28   0.02    15.07   0.00
maize                0.18   0.01    12.43   0.00
soybeans             0.31   0.04     8.59   0.00
potatoes            -0.03   0.00   -11.16   0.00
cassava              0.03   0.00    16.91   0.00
barley              -0.02   0.03    -0.61   0.54
cocoa                0.25   0.10     2.45   0.02
-----------------------------------------------

Estimated dispersion parameter = 0.06
```

This is the best fitting model using Fertilizer as one of the variables.

```
Standard errors: MLE
-----------------------------------------------
                     Est.   S.E.   t val.      p
----------------- ------- ------ -------- ------
(Intercept)         -0.58   0.17    -3.33   0.00
fertilizer           0.00   0.00     1.97   0.05
wheat                0.24   0.04     5.46   0.00
rice                -0.01   0.02    -0.56   0.58
maize                0.15   0.03     4.87   0.00
barley               0.05   0.05     0.93   0.35
soybeans             0.27   0.05     5.54   0.00
potatoes            -0.02   0.01    -3.98   0.00
beans               -0.14   0.10    -1.42   0.16
cassava              0.02   0.01     2.61   0.01
cocoa               -0.10   0.19    -0.54   0.59
bananas             -0.00   0.00    -0.50   0.62
-----------------------------------------------
```

```
Estimated dispersion parameter = 0.06
```

In this section, discuss and interpret your results. Explain the implications and significance of your findings, and relate them to your research question and existing literature or theories.

Conclusion

Throughout our analysis of the effects of different variables on the crop yield, we found that there is a high correlation between the interactions of entity, year, land, fertilizer, and the region. Year and fertilizer we found have the lowest correlation. Fertilizer having one of the lower correlations was surprising because our original thoughts were that land and fertilizer would have the greatest effects on the crop yield (both individually and for total cereal yield). However, our correlation plots did confirm our original question which was does the region and amount of arable land have a large impact on crop yield. Some factors that could also have an effect on the crop yield would be the weather of specific regions which is something our data did not provide. In future studies it would be beneficial to get weather data to look at that covariate alongside the arable land and region of where the crops are being grown and we believe that would have a more significant impact on each specific crop's yield. Based on our findings, we can say that while fertilizer can be beneficial to crop yield, the biggest factors when considering where to grow certain crops heavily depends on the region, entity, and the amount of arable land available to produce the specified crops.

References

Appendix (Optional)

If you want to include your R code as an appendix, you can create a new code chunk and set `#| echo: true` to show the code and `#| eval: false` to avoid that to be run. You can also provide scripts separately as supplementary material.

```
# Your code here
```