Exercise 9

Ian Flores, Israel Dilan

December 14, 2017

Contents

Summary

The objectives of this project were to examine the relationship between different communities in Spain and their relationship to Wheat prodcution, and to see if we could estimate a linear model that would fit this data. These analyses were done using the R language for statistical computing. The results seem to suggest that there exists a positive relationship between the size of a community and the area it dedicates to grow wheat. In the inferential component, we were able to fit a linear model that explains 86% of the variation of the data. As a conclusion we suggest the necessity of gathering similar data in other countries to be able to extrapolate as well as gathering data related to socio-economic factors that might also explain the percentage of area dedicated.

Introduction

Most countries depend, on varying degrees, on a local agricultural industry. In Spain one of the products produced is wheat. In this project we wanted to explore the relationship between different spatial features that might be influencing Spain's production and distribution of wheat. We would expect the central region of Spain to be one of the most productive ones given its harsh conditions to grow other products. Here, we make a simple Exploratory Data Analysis to explore this relationship and then proceed to see if we can make some inference regarding the area dedicated to grow wheat.

Methodology

In this project we used the R language for statistical computing. We also made use of the PASWR package for obtaining the dataset, and the ggplot2 package for visualizing the different plots. For the project, we used two DataFrames that we merged in one. The first dataset was the 'WheatSpain', containing data about seventeen spanish communities and their corresponding surface area dedicatd to growing wheat. The second dataset was the 'SurfaceSpain', containing data about the surface area for seventeen autonoumous spanish communities. Once the datasets were merged, the project was divided into two main parts.

The first part was the descriptive part. In this part we first calculated the percentage of area per community that was used to grow wheat. Then, we started by visualizing this percentage as a barplot and comparing it to a dotchart. Afterwards, we wanted to observe the relationship between the Total Surface Area and the Surface Area dedicated to growing wheat so we used a scatterplot to visualize this relationship, but also

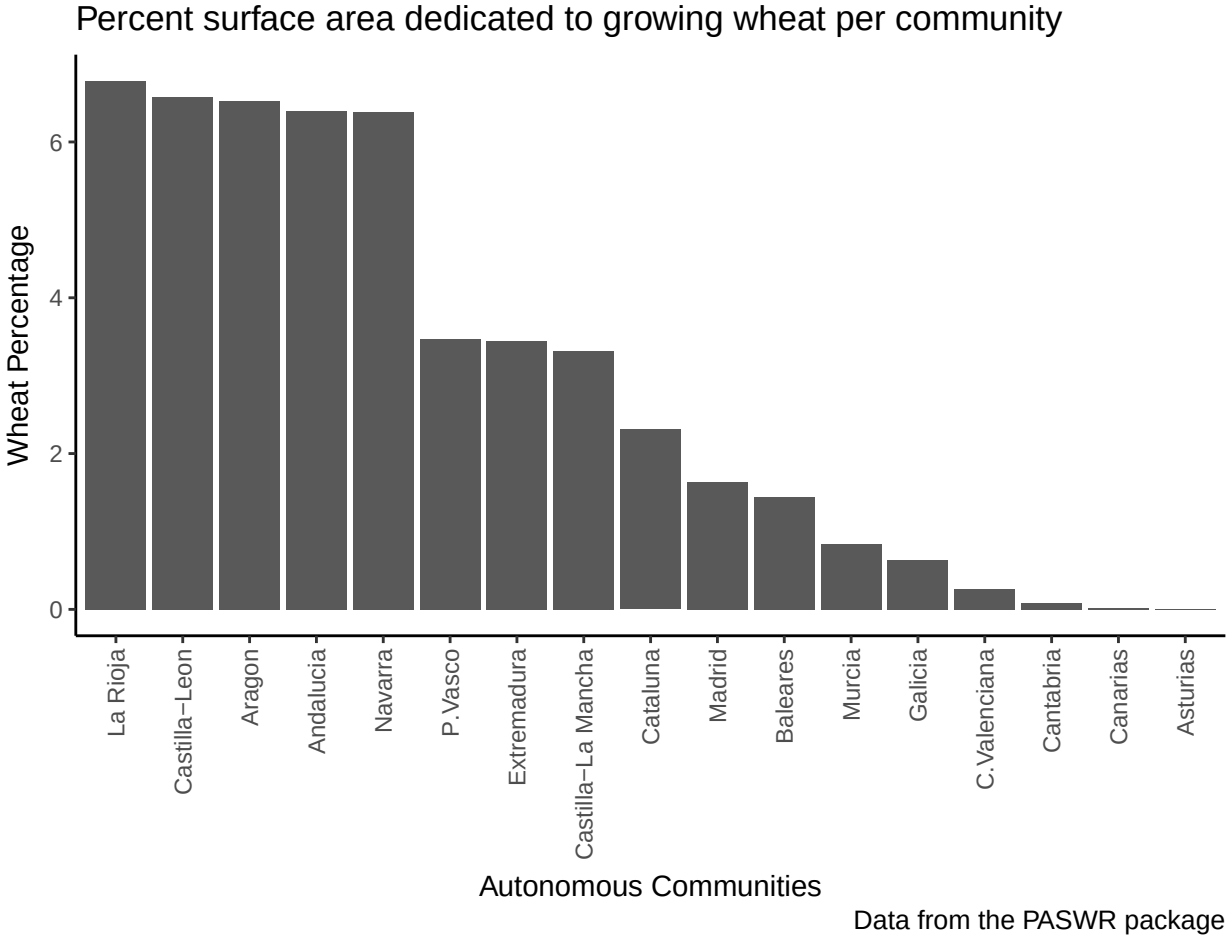
calculated the correlation between both variables to observe how strong or weak was the link. Later on, we wanted to delve into the relationship between surface area dedicated to growing wheat and the percentage of area dedicated to growing wheat. Again, we used the scatterplot as the tool to visualize this relationship and also the correlation between variables to observe how strong was the relationship.

The second part was the inferential part. In this part we developed a linear model establishing a linear relationship between the Total Surface Area and the Surface Area dedicated to growing wheat. This was done using the linear model algorithm implemented in R, which uses QR matrix decomposition as its default. Finally, we visualized this linear model over the scatterplot previously produced.

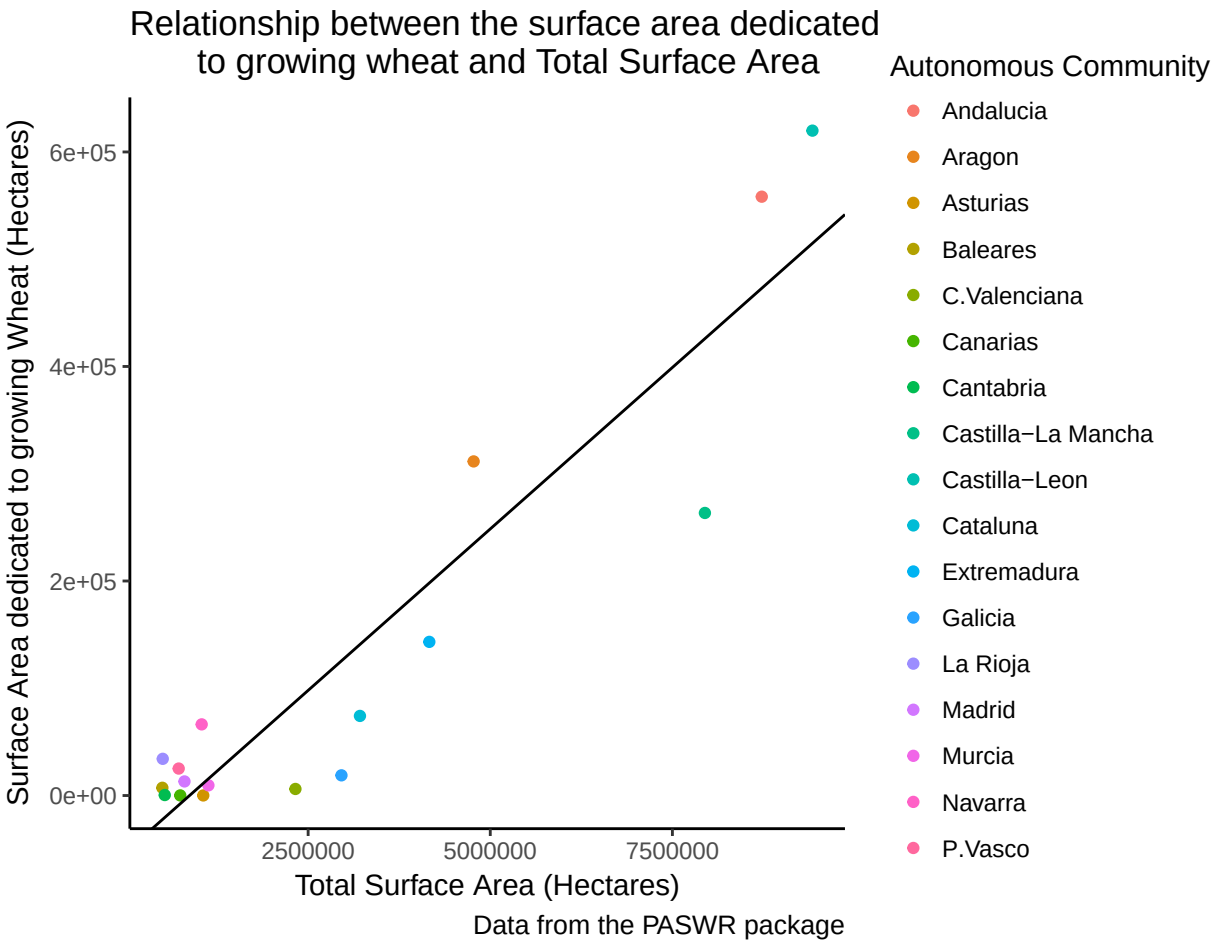
Results

Exploratory Data Analysis

In this section we were able to visualize the different relationships in the merged dataset. All of the relationships had a positive correlation.



Inferential Methods



For this part, the linear model seemed to explain aorund 86% (R^2) percent of the data which seems good enough for our pourposes here. We estimated a slope of 0.06 and an intercept of -5.253 x 10^4. This means we can observe a positive relationship between the total surface area and the surface area dedicated to growing wheat.

Conclusion

As the data suggests, from the analysis we have presented here, there is a relationship that communities who are larger tend to have a bigger percentage of area dedicated to growing wheat. This can be because of two different factors. The first one of this, is that administratively larger regions tend to be agricultural regions while smaller regions tend to be urban. The second one might be spatial aggregation of this communities. One of the limitations that we can appreciate from this project is the small dataset used. For generalization, we think it would be useful to gather more data from nearby countries to be able to extrapolate. One final suggestion would be to incorporate this data with more socio-economic data to analyze the relationships that this might have over the grow of wheat.

References

- R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- Alan T. Arnholt (2012). PASWR: PROBABILITY and STATISTICS WITH R. R package version 1.1. https://CRAN.R-project.org/package=PASWR
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Code Appendix

```
library(PASWR)
```

(a) Use the function merge() to combine the data frames WheatSpain (from problem 3) and SurfaceSpain into a new data frame named DataSpain.

```
DataSpain <- merge(WheatSpain, SurfaceSpain)
head(DataSpain)
```

```
##      community hectares     acres surface
## 1    Andalucia   558292 1379569.6   87268
## 2       Aragon   311479  769681.4   47719
## 3     Asturias       65     160.6   10604
## 4     Baleares     7203   17799.0    4992
## 5 C.Valenciana     6111   15100.6   23255
## 6     Canarias      100     247.1    7447
```

```
summary(DataSpain)
```

```
##         community     hectares          acres              surface
##  Andalucia   : 1   Min.   :    65   Min.   :    160.6   Min.   : 4992
##  Aragon      : 1   1st Qu.:  7203   1st Qu.:  17799.0   1st Qu.: 7447
##  Asturias    : 1   Median : 25143   Median :  62129.7   Median :11313
##  Baleares    : 1   Mean   :126562   Mean   : 312740.4   Mean   :29743
##  C.Valenciana: 1   3rd Qu.:143250   3rd Qu.: 353978.5   3rd Qu.:41634
##  Canarias    : 1   Max.   :619858   Max.   :1531702.5   Max.   :94223
##  (Other)     :11
```

(b) Create a variable named surface.h containing the surface area of each autonomous community in hectares. (Note: 100 hectares = 1 km 2 .) Create a variable named wheat.p containing the percent surface area in each autonomous community dedicated to growing wheat. Add the newly created variables to the data frame DataSpain and store the result as a data frame with the name DataSpain.m.

```
surface.h <- (DataSpain$surface)*100
wheat.p <- ((DataSpain$hectares)/surface.h)*100
DataSpain['Surface_Hectares'] = surface.h
DataSpain['Wheat_Percentage'] = wheat.p
```

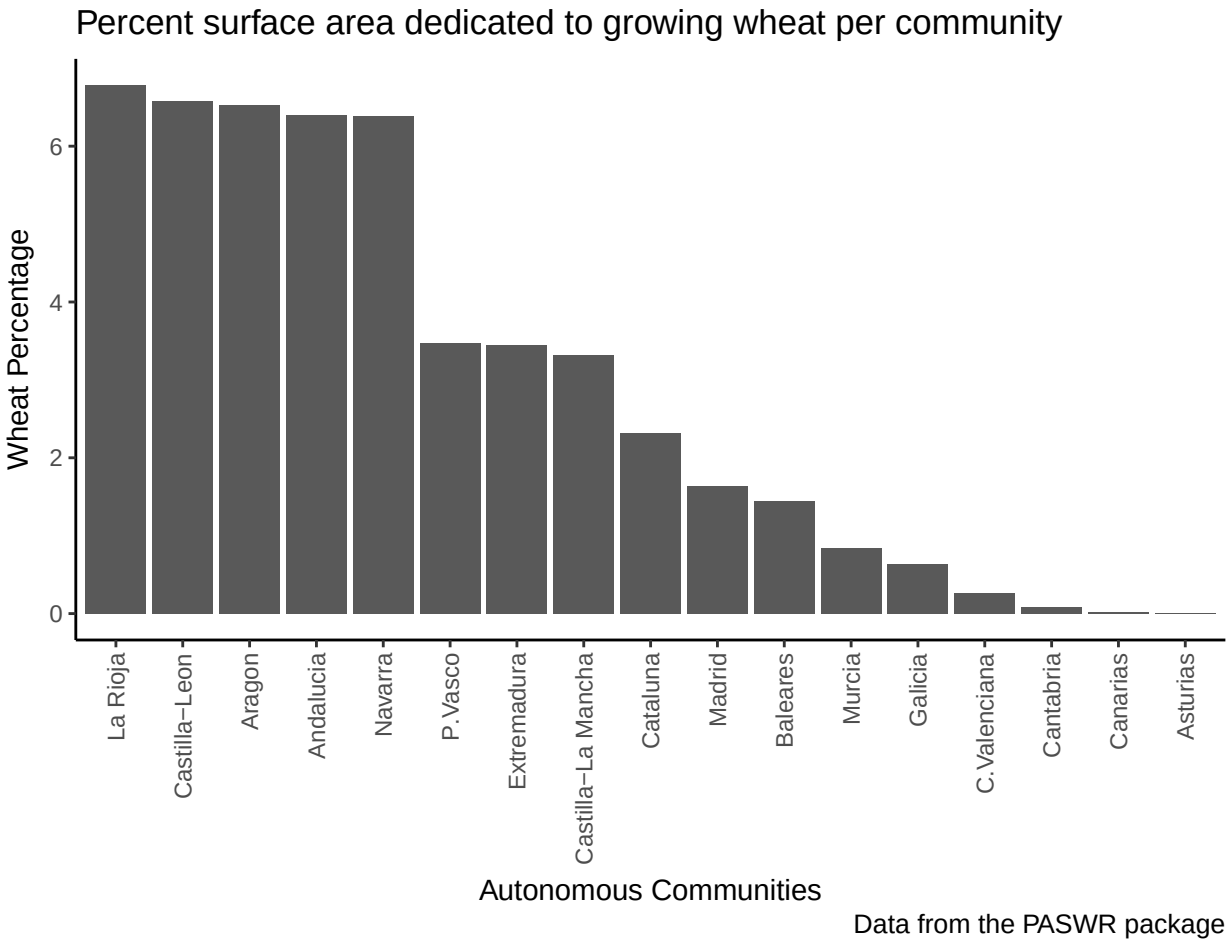
(c) Assign the names of the autonomous communities as row names for DataSpain.m and remove the variable community from the data frame.

```
rownames(DataSpain) <- DataSpain$community
DataSpain$community <- NULL
```

(d) Create a bar plot showing the percent surface area dedicated to growing wheat for each of the seventeen Spanish autonomous communities. Arrange the communities by decreasing percentages.

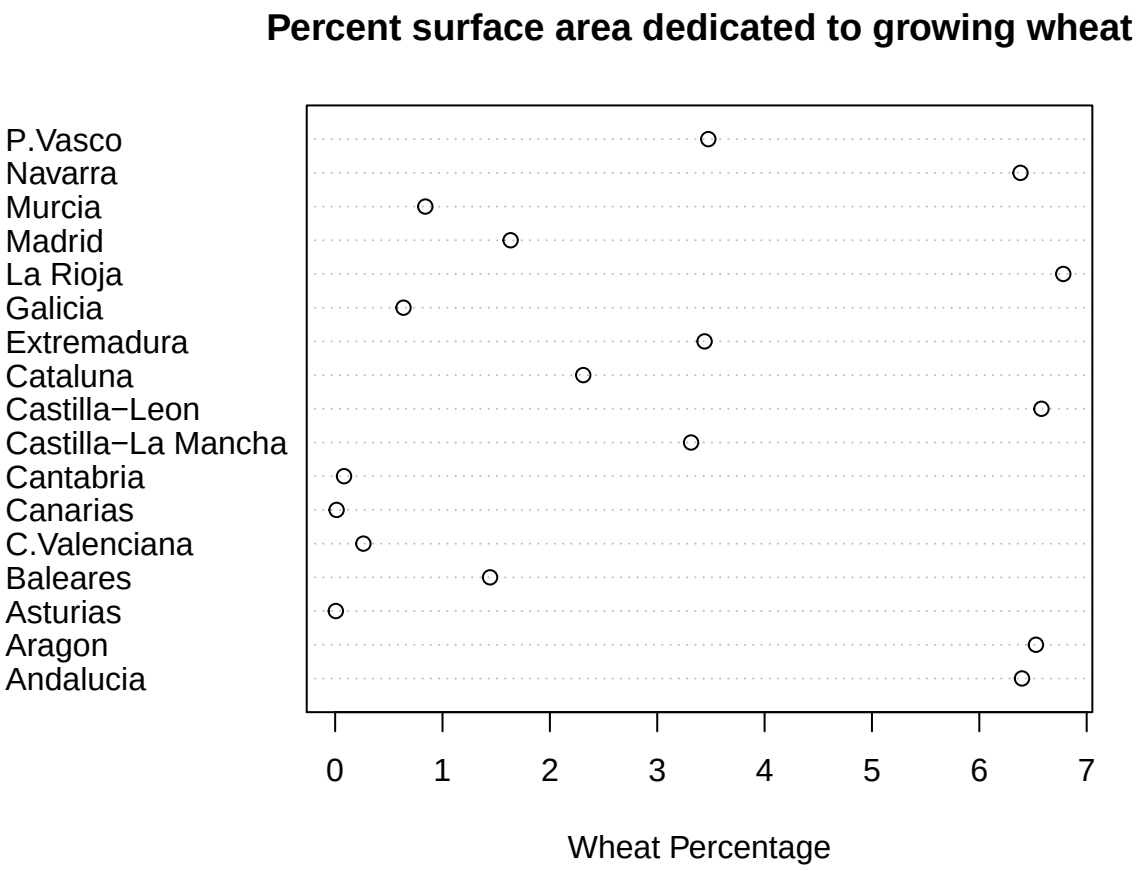
```
library(ggplot2)
ggplot(DataSpain, aes(x = reorder(rownames(DataSpain), -Wheat_Percentage),
                      y = Wheat_Percentage)) +
  geom_bar(stat = 'identity') +
```

```
  labs(x = "Autonomous Communities", y = "Wheat Percentage",
       title = "Percent surface area dedicated to growing wheat per community",
       caption = "Data from the PASWR package") +
  theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black")) +
  theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
```



(e) Display the percent surface area dedicated to growing wheat for each of the seventeen Spanish autonomous communities using the function dot chart(). To read about dot chart(), type ?dot chart at the command prompt. Do you prefer the bar chart or the dot chart? Explain your answer.

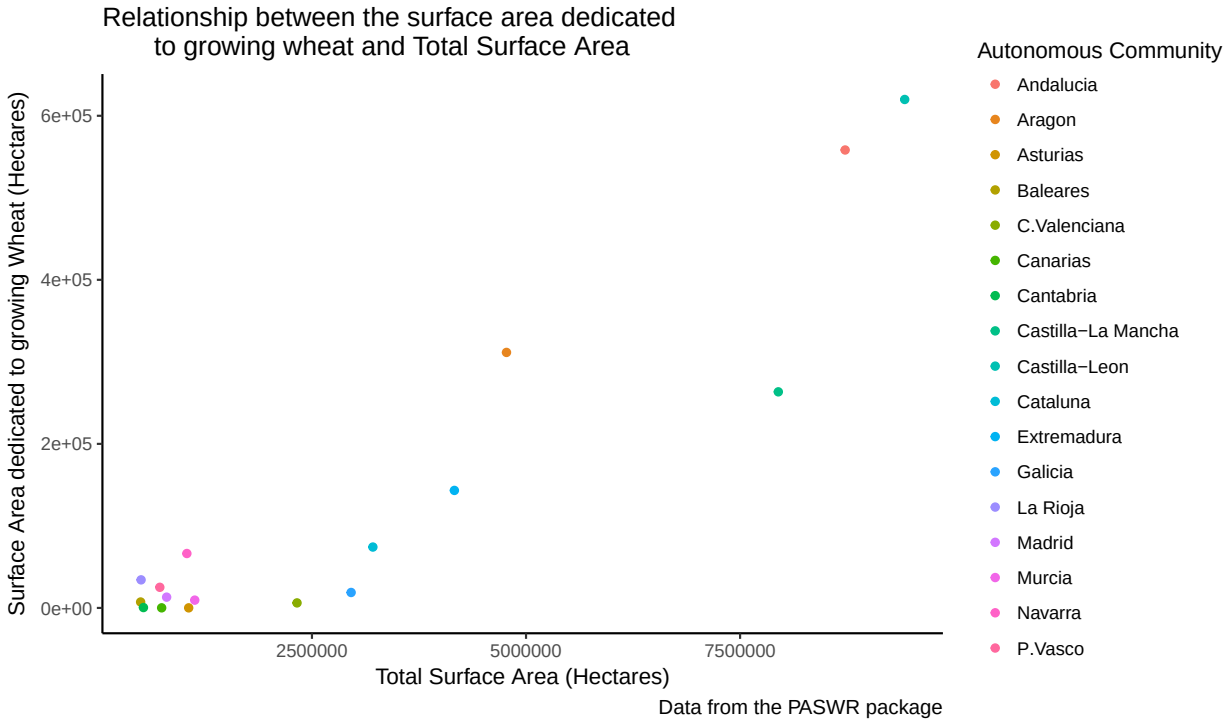
```
dotchart(DataSpain$Wheat_Percentage, labels = rownames(DataSpain),
         xlab = "Wheat Percentage",
         main = "Percent surface area dedicated to growing wheat")
```



I prefer the bar chart because it is more intuitive to follow visually than the dot chart. However, I think it can also be more misleading than the dot chart to observe differences between groups.

(f) Describe the relationship between the surface area in an autonomous community dedicated to growing wheat (hectares) and the total surface area of the autonomous community (surface.h).

```
ggplot(DataSpain, aes(x = Surface_Hectares, y =  hectares, color=rownames(DataSpain))) +
  geom_point() +
  labs(x = "Total Surface Area (Hectares)",
       y = "Surface Area dedicated to growing Wheat (Hectares)",
       title = "Relationship between the surface area dedicated
       to growing wheat and Total Surface Area",
       colour = "Autonomous Community",
       caption = "Data from the PASWR package") +
  theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```

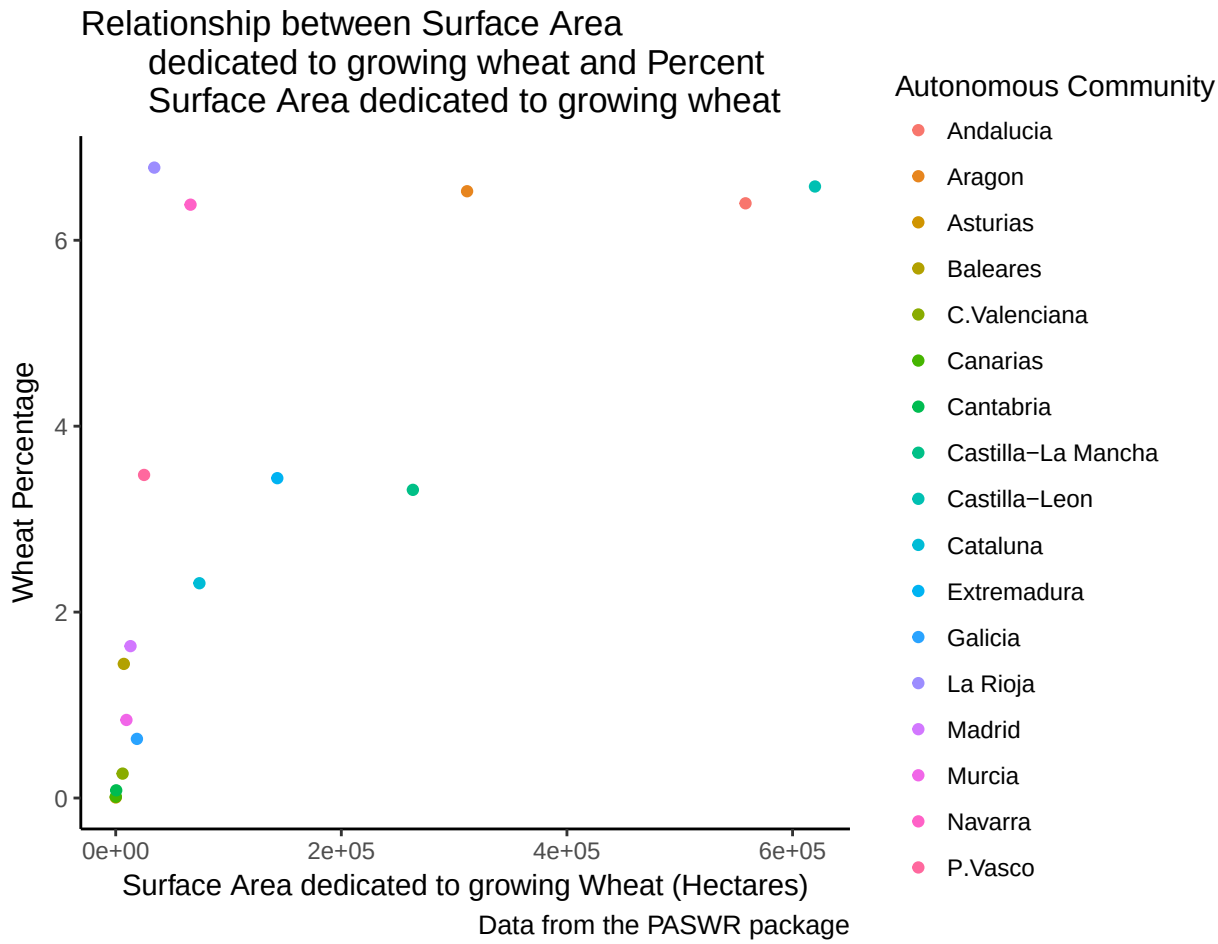



```
cor(DataSpain$Surface_Hectares, DataSpain$hectares)
```

```
## [1] 0.9289029
```

(g) Describe the relationship between the surface area in an autonomous community dedicated to growing wheat (hectares) and the percent of surface area dedicated to growing wheat out of the communities' total surface area (wheat.p).

```
ggplot(DataSpain, aes(x = hectares, y = Wheat_Percentage, color=rownames(DataSpain))) +
  geom_point() +
  labs(x = "Surface Area dedicated to growing Wheat (Hectares)",
       y = "Wheat Percentage",
       title = "Relationship between Surface Area
       dedicated to growing wheat and Percent
       Surface Area dedicated to growing wheat",
       colour = "Autonomous Community",
       caption = "Data from the PASWR package") +
  theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```




```
cor(DataSpain$hectares, DataSpain$Wheat_Percentage)
```

```
## [1] 0.6793023
```

(h) Develop a model to predict the surface area in an autonomous community dedicated to growing wheat (hectares) based on the total surface area of the autonomous community (surface.h).

```
l_m <- lm(DataSpain$hectares~surface.h)
summary(l_m)
```

```
##
## Call:
## lm(formula = DataSpain$hectares ~ surface.h)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162517  -54913   17306   56286  105043
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.253e+04  2.598e+04  -2.022   0.0614 .
## surface.h    6.021e-02  6.198e-03   9.715  7.3e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 75470 on 15 degrees of freedom
## Multiple R-squared:  0.8629, Adjusted R-squared:  0.8537
## F-statistic: 94.38 on 1 and 15 DF,  p-value: 7.302e-08
```

```
ggplot(DataSpain, aes(x = Surface_Hectares, y =  hectares, color=rownames(DataSpain))) +
  geom_point() +
 geom_abline(intercept = l_m$coefficients[1], slope = l_m$coefficients[2]) +
  labs(x = "Total Surface Area (Hectares)",
       y = "Surface Area dedicated to growing Wheat (Hectares)",
       title = "Relationship between the surface area dedicated
       to growing wheat and Total Surface Area",
       colour = "Autonomous Community",
       caption = "Data from the PASWR package") +
  theme_bw() +
  theme(panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(), axis.line = element_line(colour = "black"))
```