The Dependence of the Blue vs. Green Distinction in the Language on Certain Linguistic and Geographical Factors

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────────────────────
──── tidyverse 1.3.0 ──
```

```
## ✓ ggplot2 3.3.1     ✓ purrr   0.3.4
## ✓ tibble  3.0.1     ✓ dplyr   1.0.0
## ✓ tidyr   1.1.0     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(vcd)
```

```
## Loading required package: vcd
```

```
## Loading required package: grid
```

```
require(stats)
require(party)
```

```
## Loading required package: party
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
##
## Attaching package: 'strucchange'
```

```
## The following object is masked from 'package:stringr':
##
##     boundary
```

```
require(lsr)
```

```
## Loading required package: lsr
```

Introduction

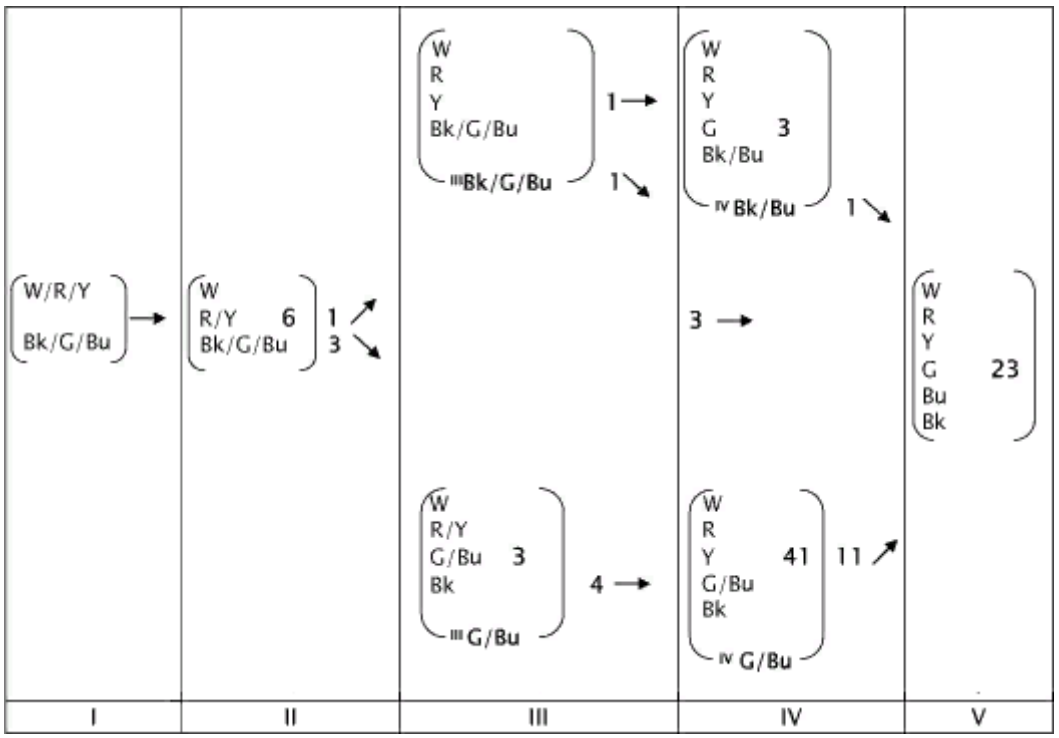
According to the universals and evolution theory of basic color terms, there are three types of colour categories:

1. Primary, there are six of them: black, white, red, yellow, green and blue.
2. Composite, which are a fuzzy unions of primariesnd are perceived by native speakers as a single color that does not consist of primary ones. Frequent example is grue (green + blue).
3. Derived, which are perceived by native speakers as a mix of primary colors (gray = black and white, purple = red and blue, etc.)

All of these color categories can be included in the basic set of colour categories in the language. Basic colours have some properties:

- have terms that are not derived from any object of this color and are not borrowed from another language (or have been borrowed for a long time and do not perceived as borrowed)
- can be combined with a wide and open set of object classes
- intuitively percieved as the primary (incl. do not consist of the terms of the other colours).

According to the evolution theory of basic color terms, languages go through several stages of development of color categories from several composites to the moment when all six primary colors receive their terms. This can happen in different ways, as shown in the image:



A Typological and Evolutionary Scheme Covering Most World Color Survey Languages (Source: Kay and Maffi 1999)

Although there are several paths, the division into blue and green always happens the last. More than that, usually derived base categories appear after stage 5, when the language already has 6 primary non-derived colour categories. But there are often exceptions.

A Hypothesis

The hypothesis of this study is that the distinction between blue and green colours in the language depends on:

- the number of basic colour categories in the language,
- red vs. yellow differentiation in the language, and possibly depends on:
- genealogical classification of the language,
- the macroarea where the language is spoken,
- geographical characteristics: longitude, latitude, altitude, and annual precipitation.

The number of non-derived basic colours seems irrelevant for this hypothesis, since it itself is rather a consequence of the presence of the blue vs. green distinction in the language.

Research Design

Variables

During the research the following variables will be used (the target variable is marked in bold):

№	name	type
1	number of non-derived colour categories in the language	numeric
2	number of base colour categories in the language	numeric

№	name	type
3	red vs. yellow differentiation in the language	binary
0	**blue vs. green differentiation in a language**	**binary**
4	genealogical classification of the language (family)	categorical
5	genealogical classification of the language (group)	categorical
6	macroarea	categorical
7	latitude	numeric
8	longitude	numeric
9	altitude	numeric
10	annual precipitation	numeric

Initially, the red vs. yellow and blue vs. green distinctions are nominal variables with a set of variants containing composites with other colours (for example, black/blue vs. green), but for this study, all variants will be reduced to binary categories. Data on the number of basic colour categories is partially fuzzy and contains intervals from the minimum to the maximum value (for example, 7-7.5). In this case, the average values will be used.

Among the available data, there are three languages in which one of the studied colours does not exist among the basic ones, so the classification is not applicable to it. It would be interesting to explore these cases, but for the purpose of this research, these objects will be deleted. In addition, the number of objects in different data differs slightly: 120 languages for chapters 132-133 and 119 languages for chapters 133-134. Information about the blue vs. green distinction and yellow vs. red distinction (or the absence of these categories) was not found for the Cahuilla language, so it will be excluded from the study.

Although the number of non-derived colours will not be used to estimate the significance of features, this variable will not be removed from the data. So, it will be possible to check whether this feature is really the consequent of the blue vs. green distinction and explains our data fully.

Since language families and groups are related features (one completely nested in the other), only one feature should be selected for the study.

Precipitation was collected from a map that was colored in the colors of the ranges of annual precipitation from maximum to minimum. The ranges are as follows: less than 100 mm, 100-250, 250-500, 500-1000, 1000-2000, 2000-3000, 3000-5000, more than 5000 mm. For each range, its minimum border was specified: a language located in an area with an annual precipitation of 500-1000 mm would receive a label of 500. Although this data is categorical in form, it is essentially numerical and will be considered so in this study.

Stages

To better understand the data, the first stage of the study will be descriptive statistics and visualization of various feature combinations.

Then statistical tests will be performed to assess the significance of the features.

Pearson’s/Spearman’s correlation test

№	Null hypothesis (H0)	Alternative hypothesis (HA)

№	Null hypothesis (H0)	Alternative hypothesis (HA)
1	There is no linear assotiation between the number of non-derived basic colours and the number of overall basic colours in the language	There is a statistically significant linear assotiation between the number of non-derived basic colours and the number of overall basic colours in the language

Based on the correlation presence, the number of overall basic color categories in the language will either be discarded, as well as non-derived, or left as a variable.

t-test

№	Null hypothesis (H0)	Alternative hypothesis (HA)
1	There is no relationship between the number of non-derived colours and the blue vs. green distinction in the language	There is a statistically significant relationship between the number of non-derived colours and the blue vs. green distinction in the language
2	There is no relationship between the number of basic colours and the blue vs. green distinction in the language	There is a statistically significant relationship between the number of basic colours and the blue vs. green distinction in the language
3	There is no relationship between the territory latitude and the blue vs. green distinction in the language	There is a statistically significant relationship between the territory latitude and the blue vs. green distinction in the language
4	There is no relationship between the territory longitude and the blue vs. green distinction in the language	There is a statistically significant relationship between the territory longitude and the blue vs. green distinction in the language
5	There is no relationship between the territory altitude and the blue vs. green distinction in the language	There is a statistically significant relationship between the territory altitude and the blue vs. green distinction in the language
6	There is no relationship between the annual precipitation and the blue vs. green distinction in the language	There is a statistically significant relationship between the annual precipitation and the blue vs. green distinction in the language

Chi-square/Fisher’s exact test, Cramer’s V, odds

№	Null hypothesis (H0)	Alternative hypothesis (HA)
1	There is no relationship between the red vs. yellow distinction and the blue vs. green distinction in the language	There is a statistically significant relationship between the red vs. yellow distinction and the blue vs. green distinction in the language
2	There is no relationship between the macroarea and the blue vs. green distinction in the language	There is a statistically significant relationship between the macroarea and the blue vs. green distinction in the language
3	There is no relationship between the language family and the blue vs. green distinction in the language	There is a statistically significant relationship between the language family and the blue vs. green distinction in the language

Two classification models will be used: decision tree and logistic regression. Models will be trained both using the non-derived variable to see how well this variable explains the data, and without it to assess the real significance of the other variables. The feature significances obtained by the two models will be compared with each other and with the results of statistical tests. In addition, the resulting decision tree will be plotted and interpreted.

It should be noted that we will not divide the data into a training and test samples, because the purpose of the study is to estimate the significance of parameters that potentially affect the green vs blue distinction in the language. We have only 116 objects, and additional separation would lead to a critical lack of data and incorrect estimates of the parameter significance.

The final stage of the research will be the generalization and interpretation of the results obtained.

The following R packages are to be used for research purposes: tidyverse, ggplot2, vcd, stats, lsr, party. Data collection and preprocessing will be performed using python.

Description of Data Collection Method

The research will be based on data obtained from the open database of structural linguistic features *World Atlas of Language Structures* (WALS), chapters 132 — 135: *Number of Non-Derived Basic Colour Categories*, *Number of Basic Colour Categories*, *Green and Blue*, *Red and Yellow*. These chapters provide the following information for languages:

1. the number of non-derived basic colour categories;
2. the number of overall basic colour categories;
3. whether red and yellow exist as two separate base colours or a composite (possibly with some third colour);
4. whether blue and green exist as two separate base colours or a composite (possibly with some third colour).

This data is the result of the World Color Survey, a project of a collaboration of researchers at the University of California, Berkeley, and the Summer Institute of Linguistics. They collected colour-naming data from 110 languages being studied by SIL translators in the late 1970s

Information about genealogy, macroarea, latitude and longitude will be also obtained from the WALS site using python. In addition, information about altitude will be collected from the Internet (https://www.advancedconverter.com/map-tools/find-altitude-by-coordinates) using python, and precipitation for all coordinates will be collected manually using specialized map (http://www.200stran.ru/maps_group28_item1439.html).

Collected Data Description

```
data <- read.csv('https://raw.githubusercontent.com/AnnaSafaryan/QAV_project/master/data.tsv'
, sep = '\t')
```

Variables:

- id
- language
- nonderived - number of non-derived color categories in the language (up to 6)
- basic - the overall number of basic color categories in the language
- gb - does the language distinguish between blue and green

- ry - does the language distinguish between red and yellow
- latitude - geographical latitude (North/South, from -90 to 90)
- longitude - geographical longitude (West/East, from -180 to 180)
- macroarea - the continent/region the language belongs to
- genus - genealogical classification (nested)
- family - genealogical classification
- altitude - the geographical altitude above sea level
- precipitation - total annual precipitation for the coordinate

Totally 116 objects without missing values were selected for the study.

```
# convert factor variables
data$gb <- as.logical(data$gb)
data$ry <- as.logical(data$ry)
data$macroarea <- as.factor(data$macroarea)
data$genus <- as.factor(data$genus)
data$family <- as.factor(data$family)
head(data)
```

```
##   id       language nonderived basic    gb    ry  latitude longitude
## 1  0         Abidji        3.5  5.00 FALSE FALSE   5.66667  -4.58333
## 2  1        Agarabi        5.0  5.00 FALSE  TRUE  -6.16667 146.00000
## 3  2 Agta (Central)        5.0  6.25 FALSE  TRUE  17.96667 121.83333
## 4  3      Aguacatec        5.0  7.25 FALSE  TRUE  15.41667 -91.33333
## 5  4     Amarakaeri        5.5  6.25 FALSE  TRUE -12.50000 -70.50000
## 6  5        Ampeeli        5.0  5.00 FALSE  TRUE  -6.75000 146.08333
##       macroarea                      genus           family altitude
## 1        Africa                        Kwa      Niger-Congo       92
## 2     Papunesia          Eastern Highlands Trans-New Guinea      974
## 3     Papunesia Greater Central Philippine     Austronesian      214
## 4 North America                      Mayan            Mayan     2609
## 5 South America                  Harakmbet        Harakmbet      257
## 6     Papunesia                      Angan Trans-New Guinea     1918
##   precipitation
## 1          1000
## 2          3000
## 3          2000
## 4          1000
## 5          1000
## 6          2000
```

Let us drop the first two columns, leaving only the features and the target variable.

```
features <- data[,3:13]
head(features)
```

```
##   nonderived basic    gb    ry  latitude longitude     macroarea
## 1        3.5  5.00 FALSE FALSE   5.66667  -4.58333        Africa
## 2        5.0  5.00 FALSE  TRUE  -6.16667 146.00000     Papunesia
## 3        5.0  6.25 FALSE  TRUE  17.96667 121.83333     Papunesia
## 4        5.0  7.25 FALSE  TRUE  15.41667 -91.33333 North America
## 5        5.5  6.25 FALSE  TRUE -12.50000 -70.50000 South America
## 6        5.0  5.00 FALSE  TRUE  -6.75000 146.08333     Papunesia
##                       genus           family altitude precipitation
## 1                       Kwa      Niger-Congo       92          1000
## 2          Eastern Highlands Trans-New Guinea      974          3000
## 3 Greater Central Philippine     Austronesian      214          2000
## 4                     Mayan            Mayan     2609          1000
## 5                  Harakmbet        Harakmbet      257          1000
## 6                     Angan Trans-New Guinea     1918          2000
```

Now let us see how many values of each category we have:

```
table(features$ry, features$gb)
```

```
##
##         FALSE TRUE
##   FALSE    15    0
##   TRUE     68   33
```

```
table(features$macroarea, features$gb)
```

```
##
##                 FALSE TRUE
##   Africa           14    2
##   Australia         1    2
##   Eurasia           5   10
##   North America    16   11
##   Papunesia        20    4
##   South America    27    4
```

```
head(table(features$family, features$gb))
```

```
##
##                FALSE TRUE
##   Afro-Asiatic     1    1
##   Algic            0    2
##   Arauan           1    0
##   Arawakan         2    1
##   Austronesian     5    2
##   Barbacoan        3    0
```

```
head(table(features$genus, features$gb))
```

```
##
##              FALSE TRUE
##   Algonquian     0    2
##   Amuzgoan       0    1
##   Angan          2    1
##   Arauan         1    0
##   Athapaskan     0    1
##   Aztecan        0    1
```

For the language family and group, there are few occurrences and a lot of zeros, so these variables do not have any predictive power, and it makes no sense to train models on them. We have to remove them from the dataset.

```
features_clean <- subset(features, select = -c(genus, family))
head(features_clean)
```

```
##   nonderived basic    gb    ry  latitude longitude     macroarea altitude
## 1        3.5  5.00 FALSE FALSE   5.66667  -4.58333        Africa       92
## 2        5.0  5.00 FALSE  TRUE  -6.16667 146.00000     Papunesia      974
## 3        5.0  6.25 FALSE  TRUE  17.96667 121.83333     Papunesia      214
## 4        5.0  7.25 FALSE  TRUE  15.41667 -91.33333 North America     2609
## 5        5.5  6.25 FALSE  TRUE -12.50000 -70.50000 South America      257
## 6        5.0  5.00 FALSE  TRUE  -6.75000 146.08333     Papunesia     1918
##   precipitation
## 1          1000
## 2          3000
## 3          2000
## 4          1000
## 5          1000
## 6          2000
```

Descriptive Statistics

```
summary(features)
```

```
##    nonderived        basic           gb              ry
##  Min.   :3.000   Min.   :3.500   Mode :logical   Mode :logical
##  1st Qu.:5.000   1st Qu.:5.000   FALSE:83        FALSE:15
##  Median :5.000   Median :6.250   TRUE :33        TRUE :101
##  Mean   :5.047   Mean   :5.901
##  3rd Qu.:5.625   3rd Qu.:7.250
##  Max.   :6.000   Max.   :9.500
##
##     latitude         longitude              macroarea
##  Min.   :-21.500   Min.   :-173.00   Africa       :16
##  1st Qu.: -5.062   1st Qu.: -77.75   Australia    : 3
##  Median :  6.208   Median : -49.17   Eurasia      :15
##  Mean   :  9.108   Mean   :  -1.94   North America:27
##  3rd Qu.: 17.438   3rd Qu.: 112.77   Papunesia    :24
##  Max.   : 67.000   Max.   : 152.77   South America:31
##
##                  genus                 family      altitude      precipitation
##  Gur                : 5   Niger-Congo     :12   Min.   :   9.0   Min.   : 250
##  Indic              : 5   Indo-European   :10   1st Qu.: 158.8   1st Qu.:1000
##  Creoles and Pidgins: 4   Trans-New Guinea: 8   Median : 307.5   Median :1000
##  Panoan             : 4   Austronesian    : 7   Mean   : 615.8   Mean   :1530
##  Angan              : 3   Oto-Manguean    : 7   3rd Qu.: 904.2   3rd Qu.:2000
##  Barbacoan          : 3   Chibchan        : 5   Max.   :2613.0   Max.   :3000
##  (Other)            :92   (Other)         :67
```

The data contains both languages that have already some derived color categories and that still contain composite ones. At the same time, most of the languages belong to a fairly narrow and relatively close to the equator zone of only about 23 degrees of latitude, so they not perfectly balanced. Also on average they belong to zones with moderate precipitation.

```
features %>%
  filter(nonderived == 6, gb == FALSE)
```

```
##   nonderived basic    gb   ry latitude longitude macroarea             genus
## 1          6  6.25 FALSE TRUE    -5.25  144.5833 Papunesia            Madang
## 2          6  6.25 FALSE TRUE    -6.25  145.6667 Papunesia Eastern Highlands
##             family altitude precipitation
## 1 Trans-New Guinea     2473          2000
## 2 Trans-New Guinea     1503          3000
```

There are only two languages in Papunesia in which there is no blue vs. green distinction, but apparently there are some other non-derived color categories. However, the number of overall basic colors is not integer, so the researchers have identified fuzzy categories.

Out of 116 languages, 44 have developed derived basic categories without developing non-derived categories for all 6 primary colours. This corresponds to the evolutionary theory of colour categories, since it allows for a significant number of exceptions.

```
features %>%
  filter(nonderived < 6, basic >= 6) -> excepts
nrow(excepts)
```

```
## [1] 44
```

```
summary(excepts)
```

```
##    nonderived        basic           gb              ry
##  Min.   :5.000   Min.   :6.250   Mode :logical   Mode :logical
##  1st Qu.:5.000   1st Qu.:6.250   FALSE:42        FALSE:1
##  Median :5.000   Median :6.250   TRUE :2         TRUE :43
##  Mean   :5.102   Mean   :6.784
##  3rd Qu.:5.000   3rd Qu.:7.250
##  Max.   :5.500   Max.   :9.500
##
##     latitude         longitude               macroarea      genus
##  Min.   :-20.000   Min.   :-173.00   Africa       : 4   Indic  : 3
##  1st Qu.: -4.146   1st Qu.: -92.48   Australia    : 1   Tacanan: 3
##  Median :  6.333   Median : -70.50   Eurasia      : 4   Angan  : 2
##  Mean   :  8.268   Mean   : -20.72   North America:13   Mayan  : 2
##  3rd Qu.: 17.804   3rd Qu.:  71.88   Papunesia    : 8   Oceanic: 2
##  Max.   : 65.000   Max.   : 152.77   South America:14   Surmic : 2
##                                                         (Other):30
##            family      altitude      precipitation
##  Oto-Manguean : 5   Min.   :  15.0   Min.   : 250
##  Austronesian : 4   1st Qu.: 148.8   1st Qu.: 500
##  Indo-European: 3   Median : 446.0   Median :1000
##  Tacanan      : 3   Mean   : 661.9   Mean   :1358
##  Uto-Aztecan  : 3   3rd Qu.: 979.2   3rd Qu.:2000
##  Chibchan     : 2   Max.   :2609.0   Max.   :3000
##  (Other)      :24
```

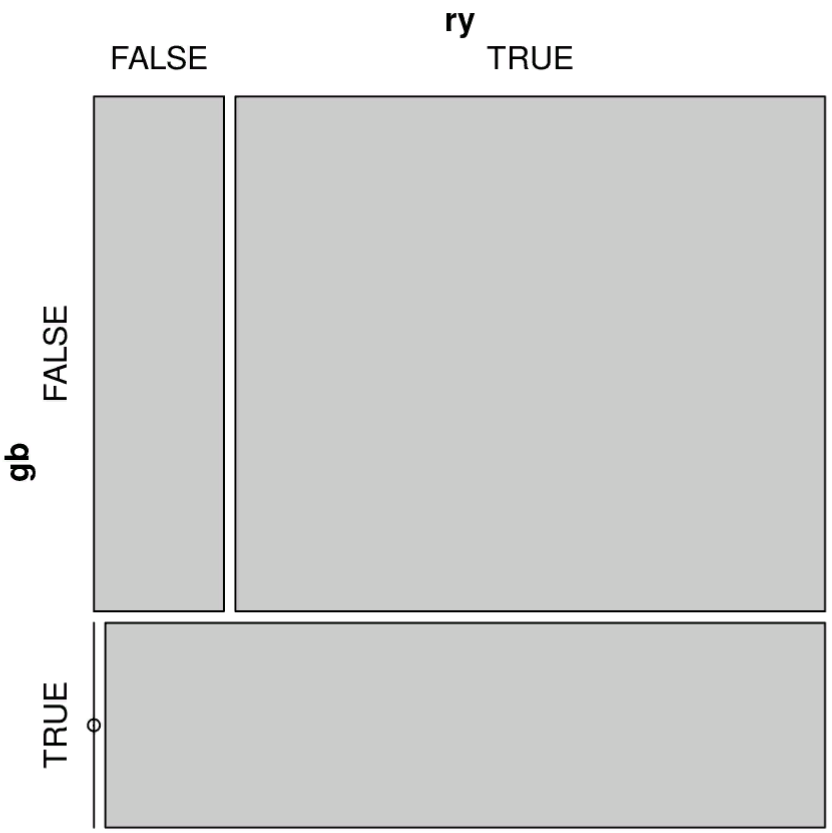
At the same time, the number of non-derived colour categories is equal to 5 or is already slightly shifted to 6, and in the vast majority of such languages there is no blue vs. green distinction, which confirms the statement that this devision occurs last. It is also interesting that most of these languages are located in North or South America at a relatively low altitude above sea level, and some are also located in an area with significant precipitation.

```
excepts %>%
  filter(precipitation == 3000)
```

```
##   nonderived basic    gb   ry latitude longitude     macroarea
## 1        5.0  6.25 FALSE TRUE -4.58333 142.83333     Papunesia
## 2        5.0  7.25 FALSE TRUE  8.00000 -77.33333 South America
## 3        5.0  6.25 FALSE TRUE -7.25000 146.00000     Papunesia
## 4        5.0  6.25 FALSE TRUE -2.75000 -71.75000 South America
## 5        5.5  6.25 FALSE TRUE -6.91667 146.58333     Papunesia
## 6        5.0  6.25 FALSE TRUE -4.00000 152.76667     Papunesia
## 7        5.0  6.25 FALSE TRUE -2.45000 140.41667     Papunesia
## 8        5.0  6.25 FALSE TRUE -4.00000 -70.50000 South America
## 9        5.0  6.25 FALSE TRUE -0.75000 -71.00000 South America
##                     genus           family altitude precipitation
## 1               Sepik Hill            Sepik       48          3000
## 2                     Kuna         Chibchan      514          3000
## 3                   Angan Trans-New Guinea     1109          3000
## 4                 Huitoto        Huitotoan      100          3000
## 5                 Oceanic     Austronesian     1332          3000
## 6                 Oceanic     Austronesian      762          3000
## 7                 Sentani          Sentani     1334          3000
## 8                  Ticuna           Ticuna       98          3000
## 9 Inland Northern Arawakan         Arawakan      162          3000
```

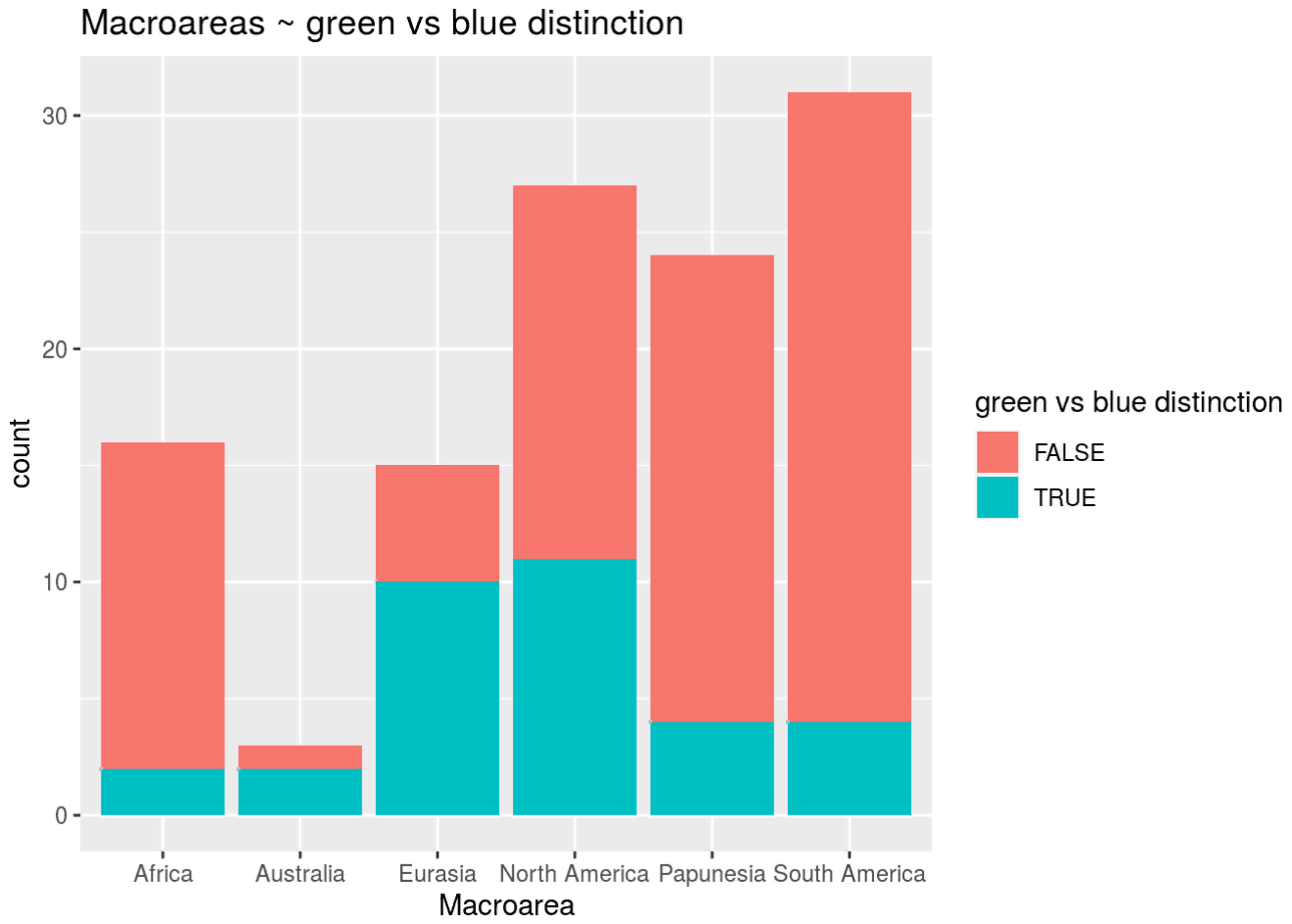
Visualisatons

```
mosaic(~ gb + ry, data=features, shade=FALSE, legend=TRUE)
```



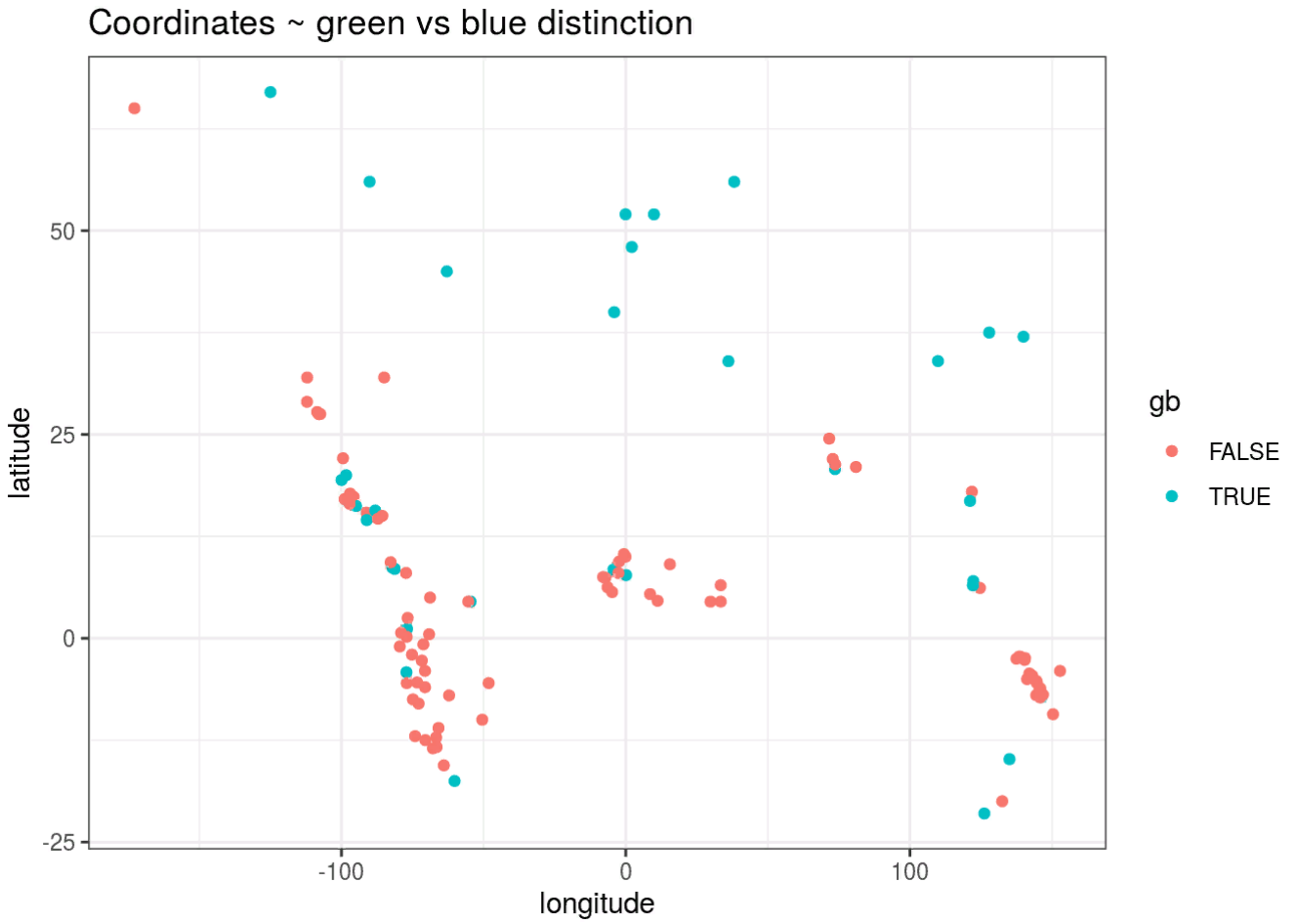
Indeed, there are no languages in the data presented that distinguish between blue and green but do not distinguish between red and yellow.

```
data %>%
  ggplot(aes(macroarea, fill=gb)) +
  geom_bar() +
  labs(title = "Macroareas ~ green vs blue distinction",
            x = "Macroarea")+
  scale_fill_discrete("green vs blue distinction")
```



Eurasia, North America and Australia have a large proportion of languages that distinguish between blue and green.

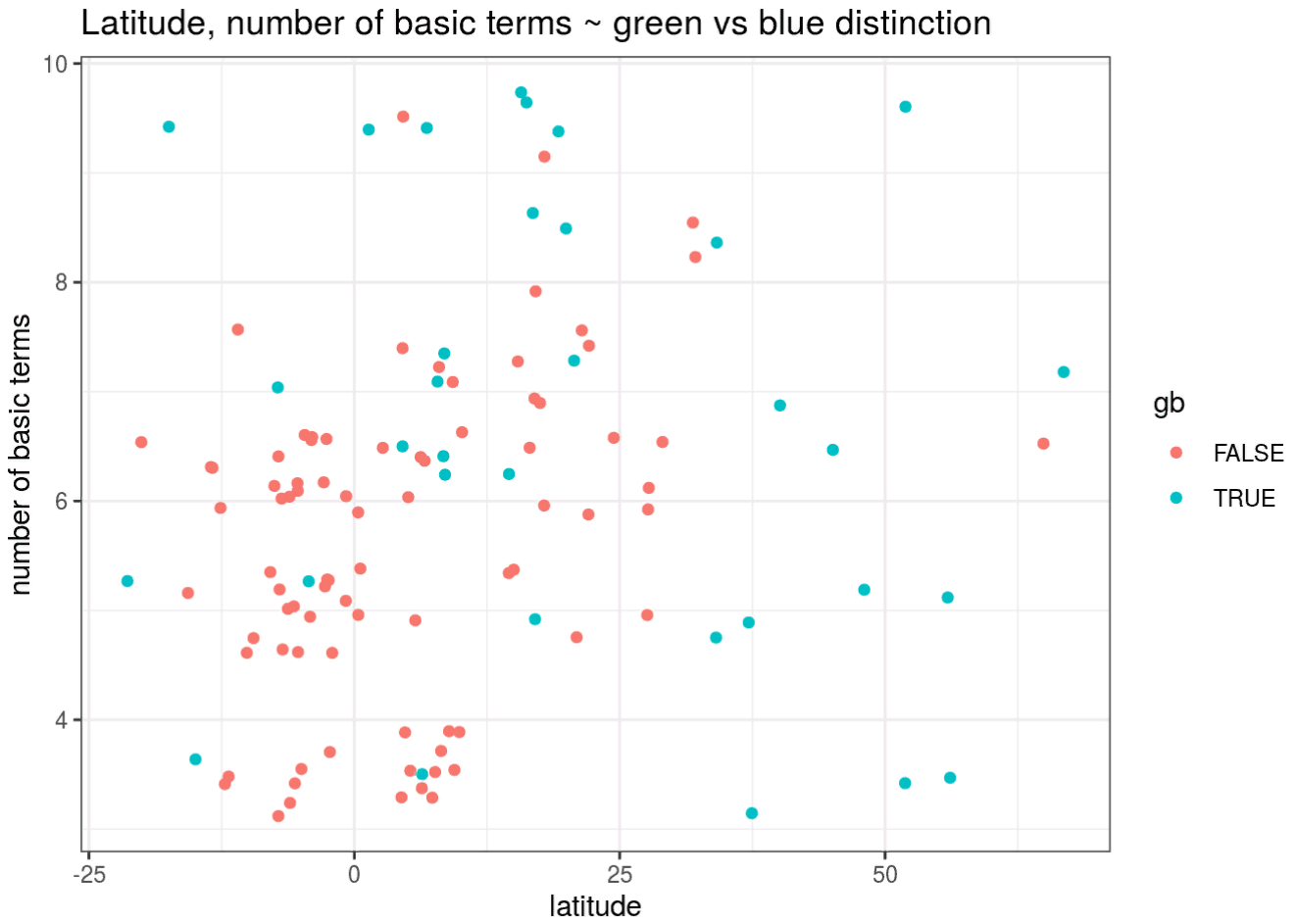
```
data %>%
  ggplot(aes(longitude, latitude, color = gb))+
  geom_jitter(width = 0.2)+
  labs(title = "Coordinates ~ green vs blue distinction",
       x = "longitude",
       y = "latitude")+
  theme_bw()
```




```
data %>%
  ggplot(aes(nonderived, basic, color = gb))+
  geom_jitter(width = 0.2)+
  labs(title = "Number of non-derived terms, number of basic terms ~ green vs blue distinctio
n",
       x = "nonderived",
       y = "basic")+
  theme_bw()
```

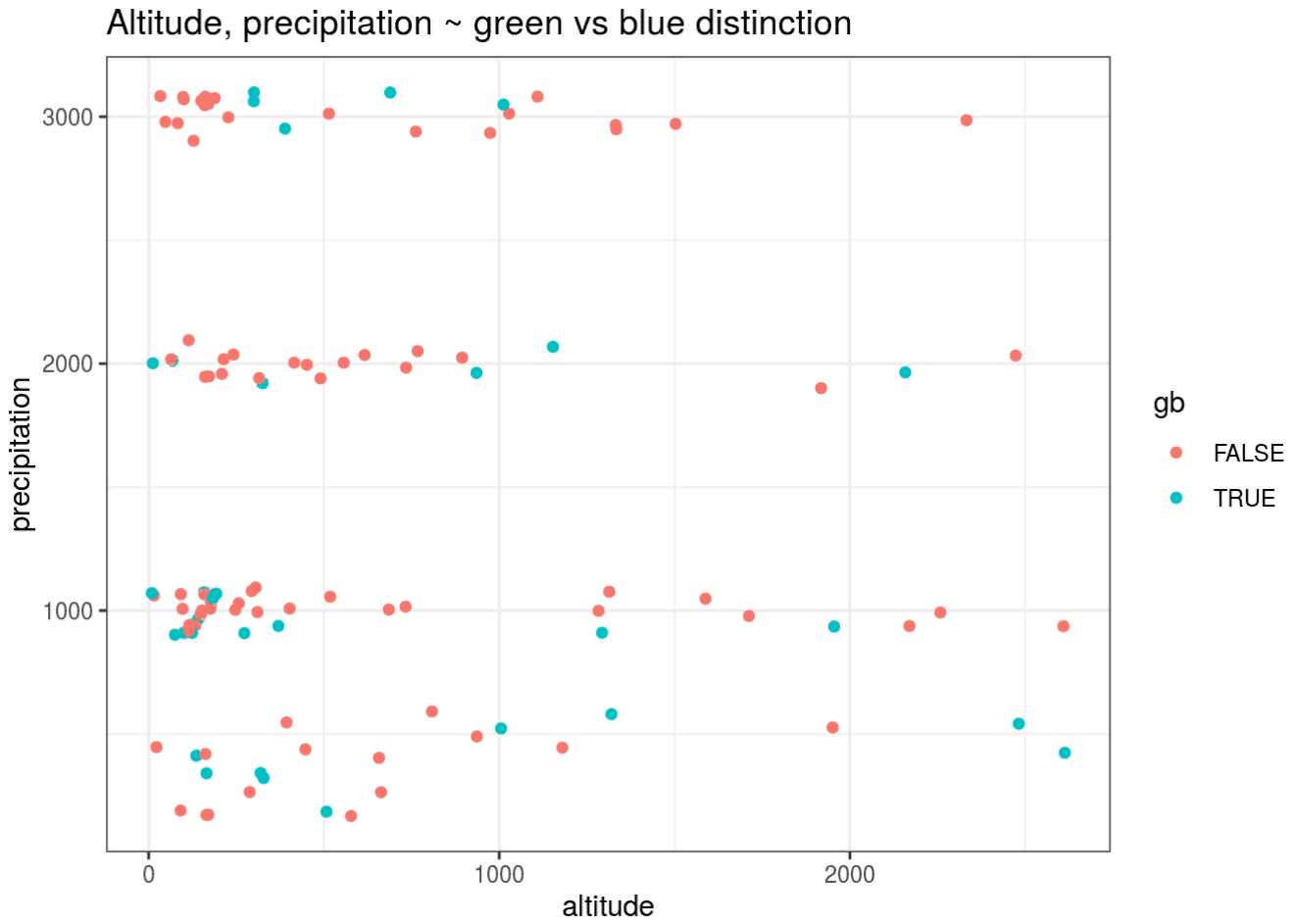
This plot already shows how well does nonderived variable explains the data.

```
data %>%
  ggplot(aes(latitude, basic, color = gb))+
  geom_jitter(width = 0.2)+
  labs(title = "Latitude, number of basic terms ~ green vs blue distinction",
       x = "latitude",
       y = "number of basic terms")+
  theme_bw()
```



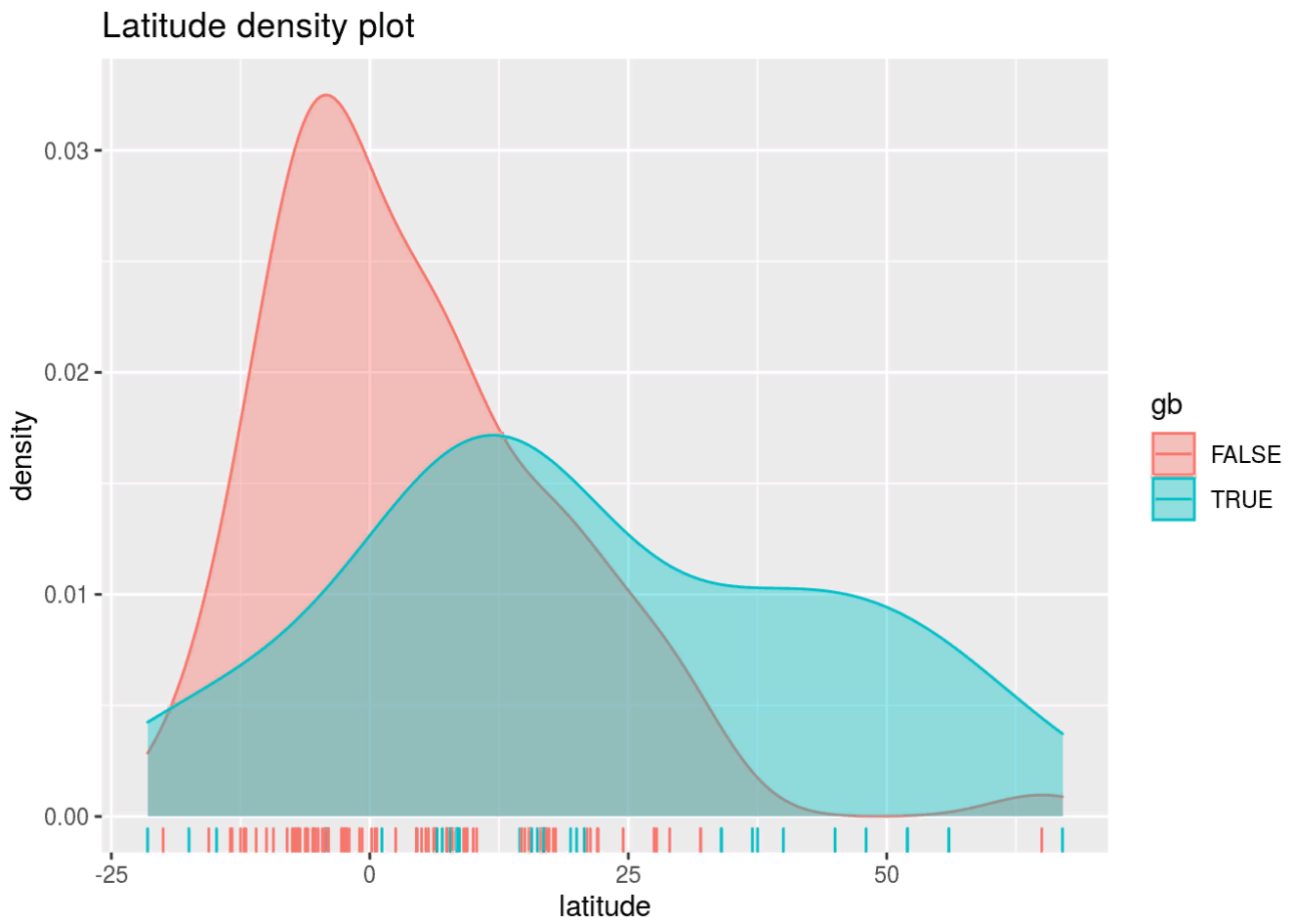
The data is better explained by latitude than by the number of basic colour categories.

```
data %>%
  ggplot(aes(altitude, precipitation, color = gb))+
  geom_jitter(width = 0.2)+
  labs(title = "Altitude, precipitation ~ green vs blue distinction",
       x = "altitude",
       y = "precipitation")+
  theme_bw()
```



Data are divided badly by altitude and precipitation.

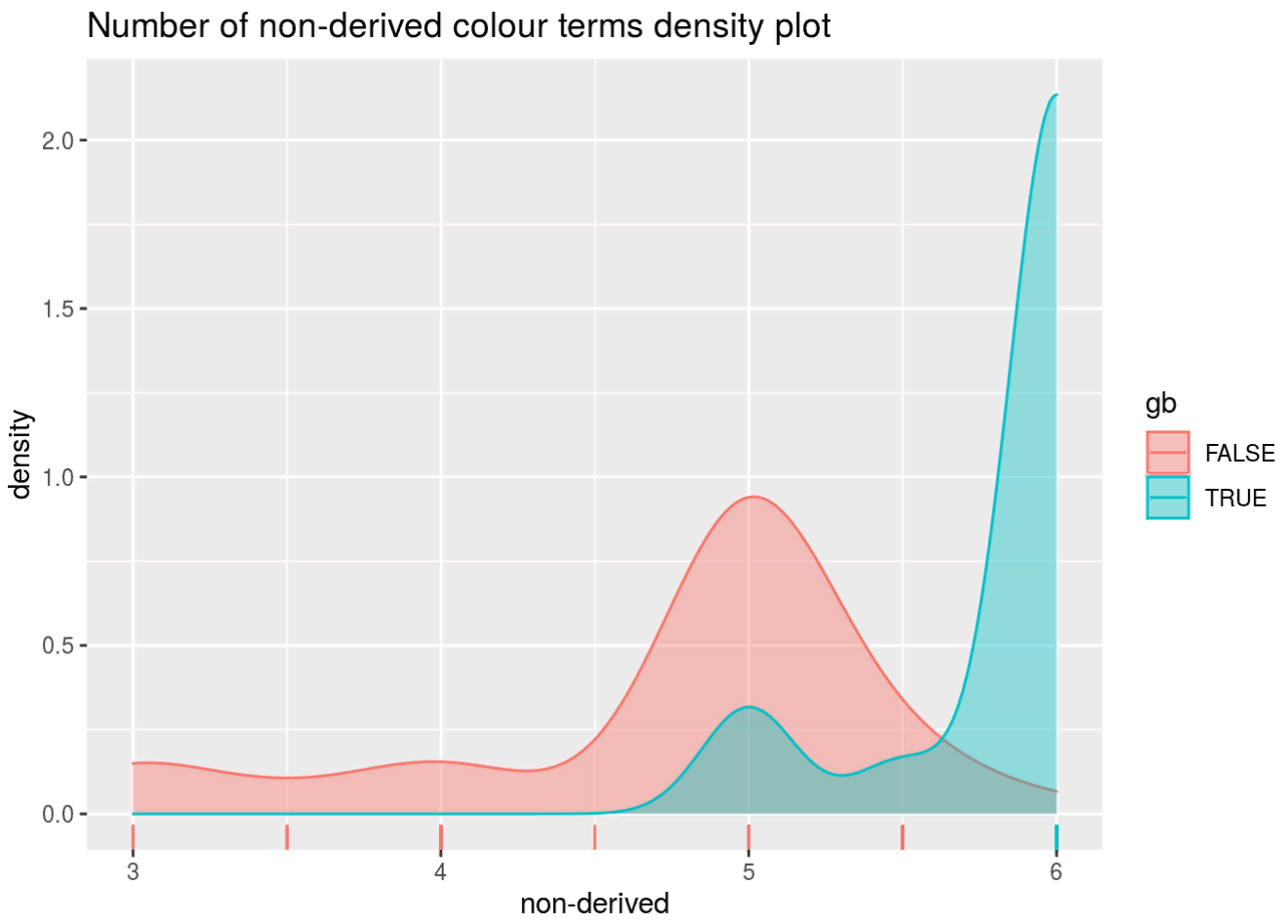
```
data %>%
  ggplot(aes(latitude, fill = gb, color = gb))+
  geom_density(alpha = 0.4)+
  geom_rug()+
  labs(title = "Latitude density plot",
       x = "latitude")
```



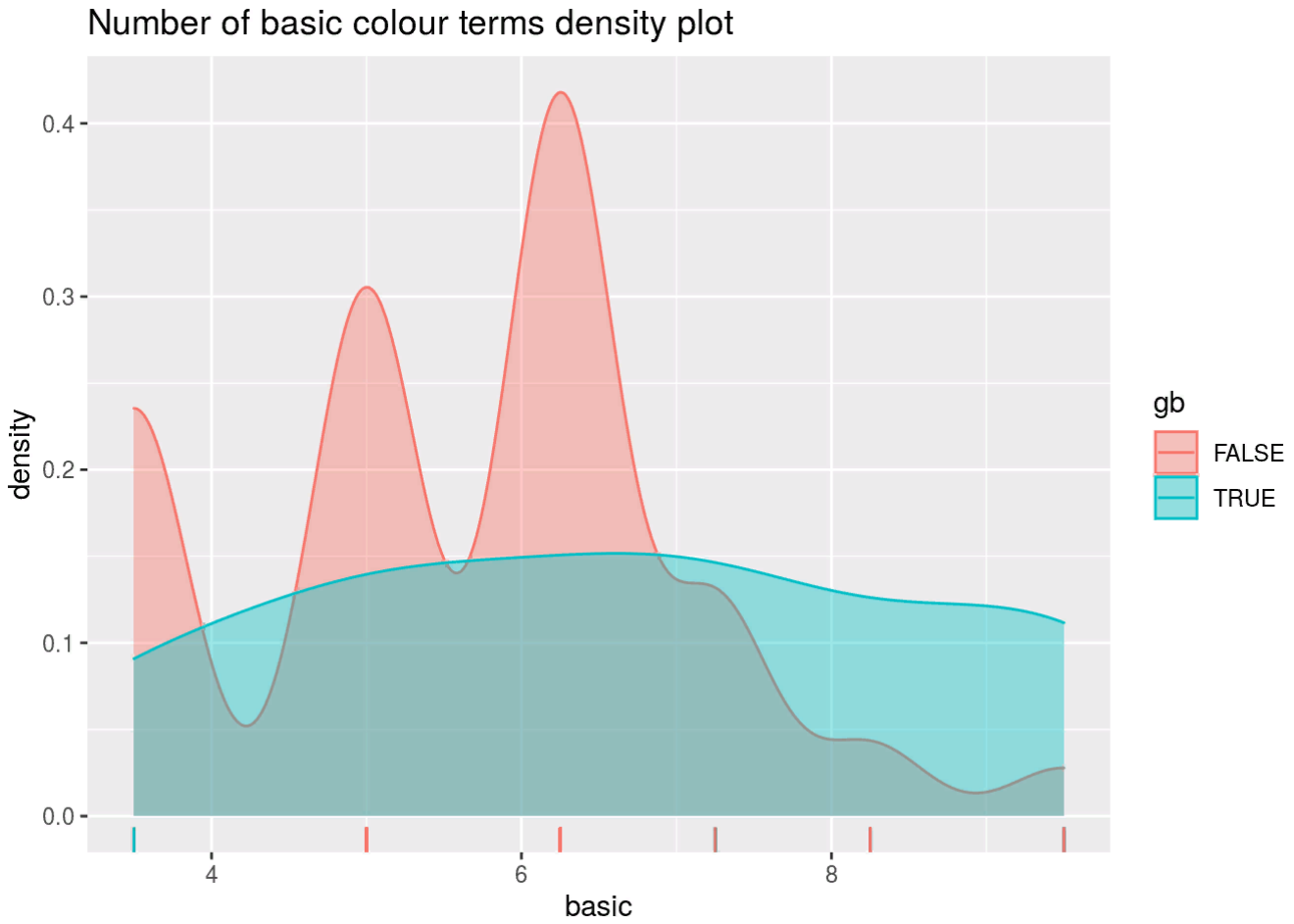
The latitude distribution for languages that distinguish between blue and green is flatter and has a higher average value. A large part of the languages without blue vs. green distinction is concentrated on lower values of latitude.

According to the evolutionary theory of colours, derived colours usually appear after all primary get separate terms, but there are many exceptions. Let us see how these variables are distributed.

```
data %>%
  ggplot(aes(nonderived, fill = gb, color = gb))+
  geom_density(alpha = 0.4)+
  geom_rug()+
  labs(title = "Number of non-derived colour terms density plot",
       x = "non-derived")
```




```
data %>%
  ggplot(aes(basic, fill = gb, color = gb))+
  geom_density(alpha = 0.4)+
  geom_rug()+
  labs(title = "Number of basic colour terms density plot",
       x = "basic")
```



The distributions are abnormal and quite different from each other.

Hypothesis Testing

Correlation for numerical variables

We can notice that the number of basic and non-derived colors have different distributions. But is there a correlation between them? Since the distribution is not normal, we use Spearman's test.

- H0: There is no linear assotiation between the number of non-derived basic colours and the number of overall basic colours in the language
- HA: There is a statistically significant linear assotiation between the number of non-derived basic colours and the number of overall basic colours in the language

```
cor.test(features$nonderived, features$basic, method = "spearman")
```

```
## Warning in cor.test.default(features$nonderived, features$basic, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  features$nonderived and features$basic
## S = 122750, p-value = 1.112e-09
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5281204
```

p-value < 0,05, so we can reject H0. There is an average correlation between the variables, but it is not possible to say that one of them linearly affects on the other completely. As already mentioned, to estimate the influence of features on the blue vs. green distinction, we can discard the non-derived variable from the data, since it is rather determined by the presence of this division, but we will not discard the overall number of basic colours.

t-tests for numerical variables

1. the number of non-derived colours

- H0: There is no relationship between the number of non-derived colours and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the number of non-derived colours and the blue vs. green distinction in the language

```
t.test(features$nonderived ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$nonderived by features$gb
## t = -11.097, df = 111.48, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3194712 -0.9196672
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.728916            5.848485
```

p-value < 0,05, so we can reject H0.

2. the number of basic colours

- H0: There is no relationship between the number of basic colours and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the number of basic colours and the blue vs. green distinction in the language

```
t.test(features$basic ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$basic by features$gb
## t = -2.5644, df = 45.063, p-value = 0.01374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7974345 -0.2160741
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            5.614458            6.621212
```

p-value < 0,05, so we can reject H0.

3. latitude

- H0: There is no relationship between the territory latitude and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the territory latitude and the blue vs. green distinction in the language

```
t.test(features$latitude ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$latitude by features$gb
## t = -3.8103, df = 41.888, p-value = 0.0004478
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24.970136  -7.677509
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.464056           20.787879
```

p-value < 0,05, so we can reject H0.

4. longitude

- H0: There is no relationship between the territory longitude and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the territory longitude and the blue vs. green distinction in the language

```
t.test(features$longitude ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$longitude by features$gb
## t = -0.18962, df = 61.569, p-value = 0.8502
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -41.97481  34.70218
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          -2.9746992           0.6616158
```

p-value > 0,05, so we can not reject H0.

5. altitude

- H0: There is no relationship between the territory altitude and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the territory altitude and the blue vs. green distinction in the language

```
t.test(features$altitude ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$altitude by features$gb
## t = -0.27827, df = 53.057, p-value = 0.7819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -334.0417  252.6456
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            604.1807            644.8788
```

p-value > 0,05, so we can not reject H0.

6. annual precipitation

- H0: There is no relationship between the annual precipitation and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the annual precipitation and the blue vs. green distinction in the language

```
t.test(features$precipitation ~ features$gb)
```

```
##
##  Welch Two Sample t-test
##
## data:  features$precipitation by features$gb
## t = 1.5619, df = 62.516, p-value = 0.1234
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -82.83852 675.39055
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1614.458            1318.182
```

p-value > 0,05, so we can not reject H0.

Chi-squares/Fisher tests for categorical variables

The Chi-square can only be applied if the expected frequency is > 5, but there are many values less in our data. We will check everything using the Fischer's exact test.

1. red vs. yellow distinction

- H0: There is no relationship between the red vs. yellow distinction and the blue vs. green distinction in the language
- H1: There is a statistically significant relationship between the red vs. yellow distinction and the blue vs. green distinction in the language

```
ry_gb <- fisher.test(features$ry, features$gb)
ry_gb
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  features$ry and features$gb
## p-value = 0.005538
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.603775      Inf
## sample estimates:
## odds ratio
##        Inf
```

```
cramersV(features$ry, features$gb)
```

```
## Warning in chisq.test(...): Chi-squared approximation may be incorrect
```

```
## [1] 0.2145256
```

p-value < 0,05, so we can reject H0. df in this case = 1, so the effect size 0.1 < 0.21 < 0.3 can be considered small.

2. macroarea

- H0: There is no relationship between the macroarea and the blue vs. green distinction in the language

- H1: There is a statistically significant relationship between the macroarea and the blue vs. green distinction in the language

```
area_gb <- fisher.test(features$macroarea, features$gb)
area_gb
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  features$macroarea and features$gb
## p-value = 0.000436
## alternative hypothesis: two.sided
```

```
cramersV(features$macroarea, features$gb)
```

```
## Warning in chisq.test(...): Chi-squared approximation may be incorrect
```

```
## [1] 0.4378314
```

p-value < 0,05, so we can reject H0. df in this case = 5, so the effect size 0.44 > 0.22can be considered large.

Results

As a result of the tests, some of the features are actually statistically significant and have impact on the presence of the blue vs. green distinction in the language.

Numeric features:

- nonderived: the average number of non-derived categories in a language that has a blue-green composite is 4.73, and in a language that distinguishes between blue and green – 5.85. It is natural that the p-value here is very small, since one event is a direct consequence of another. However, the fact that the average values in groups are not equal to integers confirms the existence of exceptions.
- basic: the average number of basic colour categories in a language that has a blue-green composite is 5.61, and in a language that distinguishes between blue and green – 6.62. However, the p-value is quite large compared to the pi-value for the nonderived variable, which confirms that it is worth including in further research.
- latitude: languages that have a blue-green composite are located souther than languages that distinguish between blue and green: the average latitude is 4.46 and 20.79 degrees of the North latitude, respectively.

Categorical features:

- ry: the presence of red vs. yellow distinction in the language has a statistically significant association with the presence of blue vs. green distinction, although the effect size is small.
- macroarea: the region has a huge effect on the distinction between blue and green in the language (however, this parameter may correlate with lalitude).

Since the family and genus parameters do not have predictive power, it does not make sense to count the Fischer’s test and Cramer’s V for them. Unfortunately, we can’t test the hypotheses related to the genealogical classification.

Models

Now let us apply classification models to our data and evaluate the significance of the parameters.

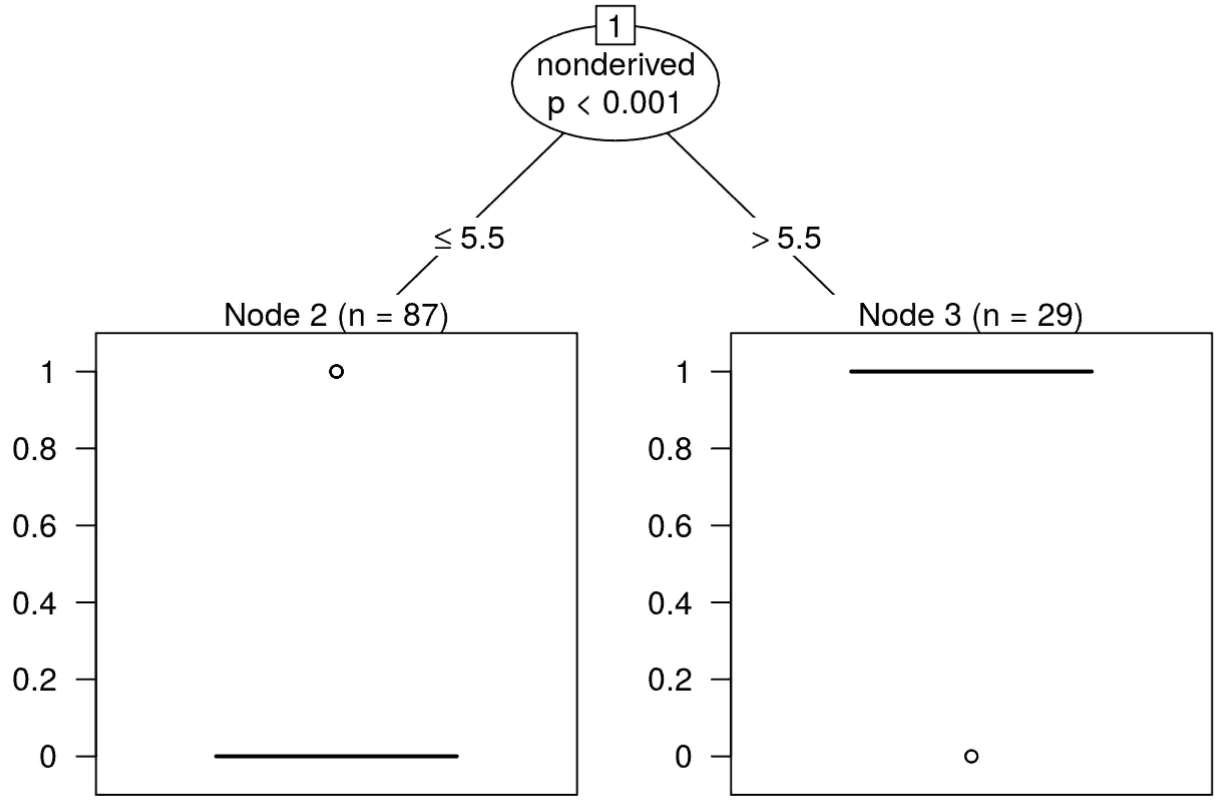
Decision Tree

Firstly train the decision tree on all the features:

```
tree_full <- ctree(gb~., data = features_clean)
print(tree_full)
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  gb
## Inputs:  nonderived, basic, ry, latitude, longitude, macroarea, altitude, precipitation
## Number of observations:  116
##
## 1) nonderived <= 5.5; criterion = 1, statistic = 43.363
##   2)*  weights = 87
## 1) nonderived > 5.5
##   3)*  weights = 29
```

```
plot(tree_full)
```



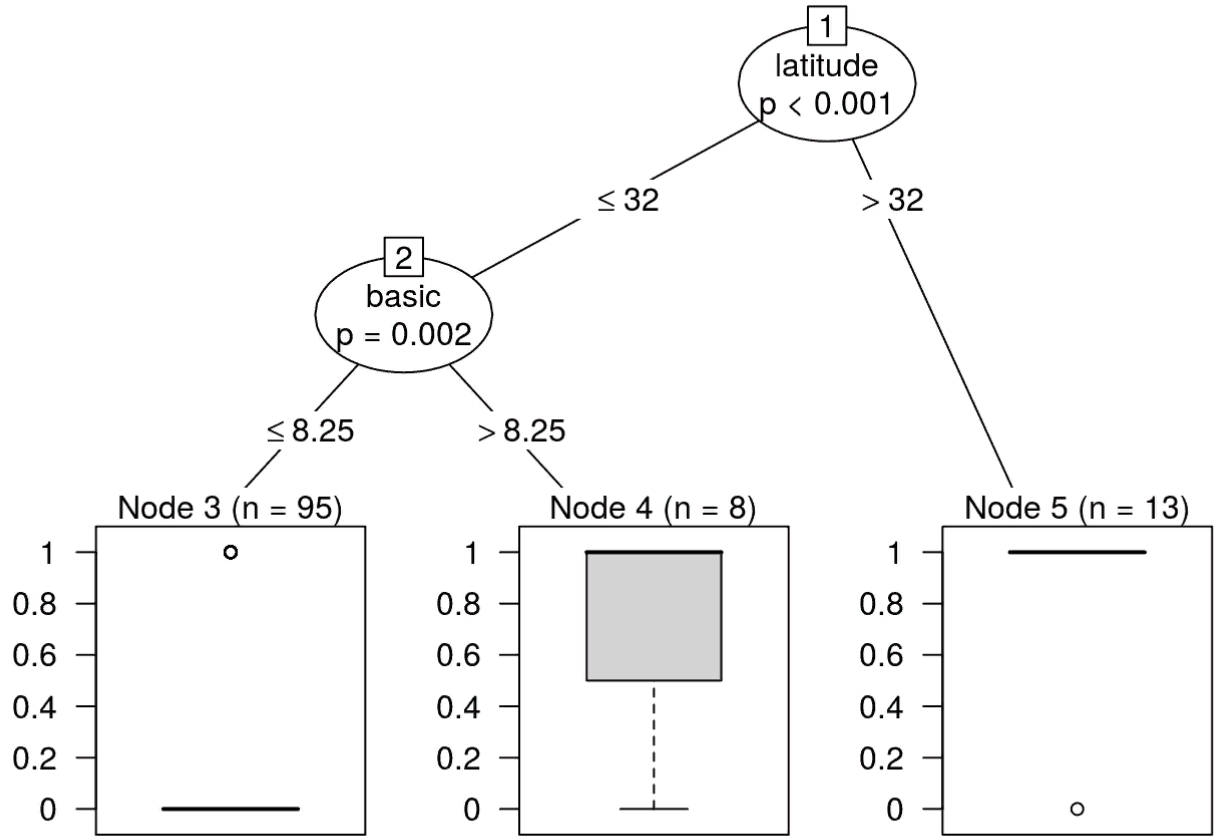
Obviously, the nonderived variable is not just statistically significant, but also almost completely explains our data. As already mentioned, this is the result of it's relations with the target variable, which confirm the World Color Survey statement: in most cases, the number of non-derived categories in a language depends directly on the wheter there is blue vs. green distinction in it or not (despite this, there are outliers in both categories that were shown during the descriptive analysis).

Now let us apply the model to the other features:

```
tree_cut <- ctree(gb~basic+ry+latitude+longitude+macroarea+altitude+precipitation, data = fea
tures_clean)
print(tree_cut)
```

```
##
##   Conditional inference tree with 3 terminal nodes
##
## Response:  gb
## Inputs:  basic, ry, latitude, longitude, macroarea, altitude, precipitation
## Number of observations:  116
##
## 1) latitude <= 32; criterion = 1, statistic = 22.045
##   2) basic <= 8.25; criterion = 0.998, statistic = 13.51
##     3)*  weights = 95
##   2) basic > 8.25
##     4)*  weights = 8
## 1) latitude > 32
##   5)*  weights = 13
```

```
plot(tree_cut)
```



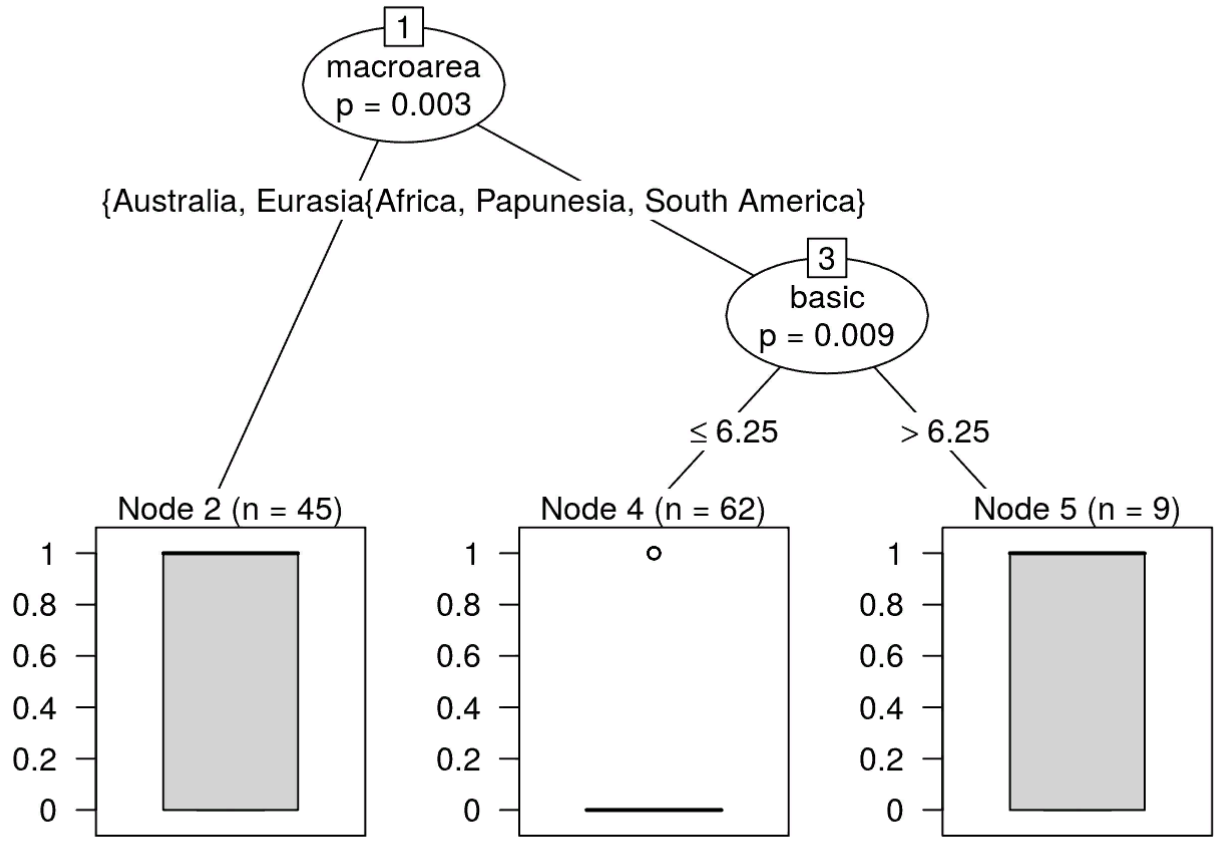
This model highlights latitude and basic as the most significant features, which coincides with the results of t-tests, which have assigned these features p-value 0.0004 and 0.014 respectively.

The fact that macroarea, which had a large effect size as a result of the Fisher's test, does not highlighted as the significant features may be explained by the relationship between this ariable and latitude. Indeed, macroarea takes the place of latitude:

```
tree_cut <- ctree(gb~basic+ry+longitude+macroarea+altitude+precipitation, data = features_cle
an)
print(tree_cut)
```

```
##
##   Conditional inference tree with 3 terminal nodes
##
## Response:  gb
## Inputs:  basic, ry, longitude, macroarea, altitude, precipitation
## Number of observations:  116
##
## 1) macroarea == {Australia, Eurasia, North America}; criterion = 0.997, statistic = 22.045
##   2)*  weights = 45
## 1) macroarea == {Africa, Papunesia, South America}
##   3) basic <= 6.25; criterion = 0.991, statistic = 10.038
##     4)*  weights = 62
##   3) basic > 6.25
##     5)*  weights = 9
```

```
plot(tree_cut)
```



But removing the number of basic colours does not put any other feature in the second place.

```
tree_cut <- ctree(gb~ry+latitude+longitude+macroarea+altitude+precipitation, data = features_
clean)
print(tree_cut)
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  gb
## Inputs:  ry, latitude, longitude, macroarea, altitude, precipitation
## Number of observations:  116
##
## 1) latitude <= 32; criterion = 1, statistic = 22.045
##   2)*  weights = 103
## 1) latitude > 32
##   3)*  weights = 13
```

Results

Applying the decision tree to the features not only confirmed the significance of the features highlighted by t-test and the Fisher's test, but also provided additional information about the threshold values:

- Languages located above 32 degrees of the North latitude tends to distinguish between blue and green.
- Languages located below 32 degrees of the North latitude may not have the blue vs. green distinction, although the number of basic colours is close to 8 (that is, they are exceptions to the postulate that derived colors appear after the primary colors gets their non-derived categories).
- Macroarea does not allow us to divide languages as clearly as latitude, but there is also something interesting: despite the fact that Australia is located below 32 degrees of the North latitude, it is in the same group with such continents as North America and Eurasia.

Logistic Regression

It is interesting to compare the results of the decision tree with the results of logistic regression.

As in the previous paragraph, we train logistic regression on all parameters at first.

```
logreg_full <- glm(gb~., data = features_clean)
summary(logreg_full)
```

```
##
## Call:
## glm(formula = gb ~ ., data = features_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6180  -0.1983  -0.0843   0.1840   0.9199
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.418e+00  2.122e-01  -6.684 1.22e-09 ***
## nonderived              4.217e-01  6.367e-02   6.624 1.63e-09 ***
## basic                  -1.655e-02  2.537e-02  -0.653   0.5155
## ryTRUE                 -2.651e-01  1.328e-01  -1.997   0.0485 *
## latitude                8.893e-03  3.856e-03   2.306   0.0231 *
## longitude               1.589e-03  1.268e-03   1.254   0.2128
## macroareaAustralia      1.494e-01  2.643e-01   0.565   0.5731
## macroareaEurasia       -3.353e-01  1.926e-01  -1.741   0.0846 .
## macroareaNorth America -1.191e-01  1.765e-01  -0.675   0.5013
## macroareaPapunesia     -3.708e-01  2.042e-01  -1.816   0.0723 .
## macroareaSouth America -5.267e-02  1.788e-01  -0.295   0.7689
## altitude               -7.223e-06  5.489e-05  -0.132   0.8956
## precipitation          -8.256e-06  4.390e-05  -0.188   0.8512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1089066)
##
##     Null deviance: 23.612  on 115  degrees of freedom
## Residual deviance: 11.217  on 103  degrees of freedom
## AIC: 86.203
##
## Number of Fisher Scoring iterations: 2
```

```
exp(logreg_full$coefficients)
```

```
##            (Intercept)             nonderived                  basic
##              0.2421062              1.5245691              0.9835847
##                  ryTRUE               latitude              longitude
##              0.7671033              1.0089330              1.0015907
##     macroareaAustralia       macroareaEurasia macroareaNorth America
##              1.1611664              0.7151317              0.8877281
##     macroareaPapunesia macroareaSouth America               altitude
##              0.6901635              0.9486895              0.9999928
##          precipitation
##              0.9999917
```

As with the decision tree, the most important features are again nonderived and latitude. Interestingly, that the macroarea is only meaningful when the language belongs to Eurasia or Papunesia. Among the features that turned out to be significant as a result of statistical tests, the significance of the presence of red vs. yellow distinction in the language was also confirmed, but the number of basic colour categories did not highlighted as significant – it was probably outweighed by the number of non-derived colours. In general, nonderived received the highest coefficient and significance code, which again coincides with the World Colour Survey statement that it is a consequence of our target variable.

Now let us try to remove the nonderived variable and see what factors affect the distinction between blue and green (since the previous results are biased due to the presence of nonderived variable).

```
logreg_clean <- glm(gb~.-nonderived, data = features_clean)
summary(logreg_clean)
```

```
##
## Call:
## glm(formula = gb ~ . - nonderived, data = features_clean)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70730  -0.25511  -0.08036   0.23186   0.96372
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -4.083e-01  1.754e-01  -2.329 0.021816 *
## basic                   6.188e-02  2.666e-02   2.321 0.022214 *
## ryTRUE                  9.951e-02  1.436e-01   0.693 0.489889
## latitude                1.527e-02  4.437e-03   3.442 0.000833 ***
## longitude               2.999e-03  1.485e-03   2.019 0.046033 *
## macroareaAustralia      5.111e-01  3.074e-01   1.663 0.099338 .
## macroareaEurasia       -1.670e-01  2.268e-01  -0.736 0.463270
## macroareaNorth America  1.038e-01  2.059e-01   0.504 0.615022
## macroareaPapunesia     -4.355e-01  2.424e-01  -1.796 0.075331 .
## macroareaSouth America  2.711e-01  2.044e-01   1.326 0.187591
## altitude                4.752e-05  6.448e-05   0.737 0.462773
## precipitation           5.249e-05  5.102e-05   1.029 0.305953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1538028)
##
##     Null deviance: 23.612  on 115  degrees of freedom
## Residual deviance: 15.995  on 104  degrees of freedom
## AIC: 125.36
##
## Number of Fisher Scoring iterations: 2
```

```
exp(logreg_clean$coefficients)
```

```
##            (Intercept)                   basic                  ryTRUE
##              0.6647639               1.0638398               1.1046272
##               latitude               longitude      macroareaAustralia
##              1.0153893               1.0030035               1.6671280
##       macroareaEurasia macroareaNorth America      macroareaPapunesia
##              0.8462037               1.1094269               0.6469301
## macroareaSouth America                altitude           precipitation
##              1.3113939               1.0000475               1.0000525
```

As in the case of the decision tree, latitude is the most important parameter here. Then, as with statistical tests, the number of basic colour categories and partially macroarea are highlighted. Interestingly, longitude is also highlighted, although t-test did not reveal its effect on the target variable, but the distinction between red and

yellow, which is significant according to the Fisher's test, is not highlighted. Australia became significant again (its special position was also shown by the decision tree) and Papunesia remained significant (this is explained by the results of the previous descriptive analysis).

Despite the lower number of parameters compared to the first model, the Akaike criterion has become higher, so this model approximates data worse. This is understandable because we have removed the variable that directly follows from the target one.

Based on the information obtained, we can build a logistic regression model that approximates our data in the best way with the smallest number of variables according to the AIC. All experiments and their resilts are here (https://github.com/AnnaSafaryan/QAV_project/blob/master/aic.csv)

```
logreg_small <- glm(gb~basic+latitude+longitude+macroarea, data = features_clean)
summary(logreg_small)
```

```
##
## Call:
## glm(formula = gb ~ basic + latitude + longitude + macroarea,
##     data = features_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7427  -0.2278  -0.0863   0.1688   0.9413
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.310351   0.154563  -2.008 0.047171 *
## basic                   0.067668   0.024419   2.771 0.006590 **
## latitude                0.014771   0.004277   3.454 0.000793 ***
## longitude               0.003161   0.001449   2.181 0.031335 *
## macroareaAustralia      0.507302   0.295118   1.719 0.088509 .
## macroareaEurasia       -0.123335   0.219476  -0.562 0.575325
## macroareaNorth America  0.193894   0.189583   1.023 0.308739
## macroareaPapunesia     -0.313517   0.220885  -1.419 0.158700
## macroareaSouth America  0.335616   0.185438   1.810 0.073125 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1518814)
##
##     Null deviance: 23.612  on 115  degrees of freedom
## Residual deviance: 16.251  on 107  degrees of freedom
## AIC: 121.21
##
## Number of Fisher Scoring iterations: 2
```

Results and Discussion

The results of the study can be summarized as follows:

1. The distinction between blue and green in a language is more influenced by its geographical position, and less by whether the language distinguishes between red and yellow.
2. Languages located closer to the North (above 32 degrees of the North latitude) are more likely to distinguish between blue and green than languages located souther.

3. Languages closer to the South are more likely to distinguish between blue and green if they already have more than 8 basic colour categories.
4. The languages of Australia behave more like the languages of North America and Eurasia than the languages of South America, Africa, or even neighboring Papunesia. Perhaps this is a result of more dense communication with English and other European languages, since most of the continent is uninhabitable, and people have to live closely.
5. In Papunesia, there are two genealogically close and neighboring languages in which blue and green have not yet got their separate categories, but some other non-derived ones have been added to the basic set.
6. Many languages that have developed derived basic categories without developing non-derived ones for all 6 primary colours are located in North or South America at a relatively low altitude above sea level (although the latter was not highlighted as a significant feature).
7. There is no connection between the blue vs. green distinction and the total annual precipitation or the territory altitude.

Perhaps these results are due to the fact that the languages of Eurasia and North America are more influenced by the world languages (English, Chineese, Russian, etc.) in which this division is present. In addition, the result may be affected by the colour of the water: in south countries with a lot of sun and high temperature, it seems greener due to the large number of algae and micro-organisms living in it.

It was impossible to estimate the significance of genealogical features, since the number of examples in each category was too small. For further research, it would be useful to group families on some basis (but not geographical). In addition, it may be a good idea to use data such as the average temperature and type of territory, and the presence of the sea nearby (this study used altitude, but it does not always mean the same).

In addition, our data is almost entirely explained by the number of non-derived base color categories (a threshold value of 5.5). This corresponds to the statement of the evolutionary theory of colour that the division into blue and green happens last. We also found enough exceptions to the postulate that languages usually develop derived basic categories only after stage 5, when all primary colours have got separate terms (evolutionary theory allows these exceptions). This may explain the small dependence of the target variable on the distinction between red and yellow.

References

- Hardin, C. L. and Maffi, Luisa (eds.) 1997. Color Categories in Thought and Language. Cambridge: Cambridge University Press.
- Kay, Paul and Berlin, Brent and Merrifield, William. 1991. Biocultural Implications of Systems of Color Naming. Journal of Linguistic Anthropology 1. 12-25.
- Kay, Paul and Berlin, Brent. 1997. Science ≠ Imperialism: There are non-trivial constraints on color naming. Behavioral and Brain Sciences 20. 196-201. (Peer Commentary on Saunders and van Brakel).
- Kay, Paul and Berlin, Brent and Maffi, Luisa and Merrifield, William. 1997. Color Naming across Languages. In Hardin, C. L. and Maffi, Luisa (eds.), Color Categories in Thought and Language, 21-56. Cambridge: Cambridge University Press.
- Kay, Paul and Maffi, Luisa. 1999. Color Appearance and the Emergence and Evolution of Basic Color Lexicons. American Anthropologist 101. 743-760.
- WALS (https://wals.info)