ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 2

Model Solutions

[Note: all data in this homework are simulated.]

[The plots included in these model solutions are fairly rough to reflect the type of plots needed for preliminary data exploration. If you need to write a report on your data exploration process, the plots would need to be tidied up.]

Standard Questions

1. *The file* `HW2Q1.txt` *contains the following data*

Variable	*Meaning*
population	*The population of the district.*
average.income	*The average annual income of full-time workers aged 18–65*
income.inequality	*A measure of inequality in income in the region with 0 representing perfect equality and 100 representing maximal inequality.*
percent.unemployed	*The percentage of people aged 18–65 unemployed in the region.*
education.level	*The average number of years spent in full-time education*
police.officers	*Number of police officers per 100,000 inhabitants*
government.spending	*Amount of government spending on the district per inhabitant.*
crime.rate	*Number of crimes committed per 1,000,000 inhabitants.*

These data were collected from government data in a variety of countries. Government spending, police officers and crime rate data were from government reports. Income data were from government tax data. Unemployment data were from government records of individuals claiming unemployment benefits. Population data is from the government census.

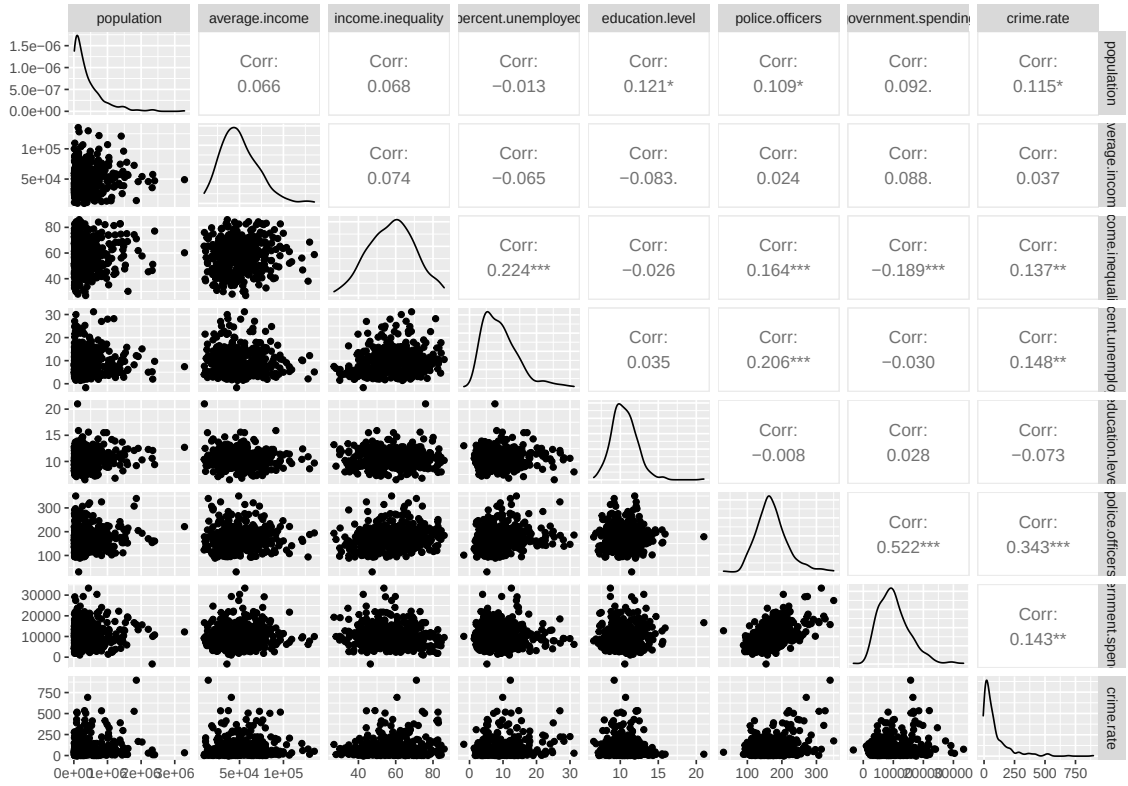
Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

We start by considering the data sources. Government data sources might be fairly reliable, depending on the government. There may be bias if the government is able to alter the figures to make themselves look better. There may also be discrepancies between the way the data is obtained in different countries. For example, if the income data is obtained from tax

records, then different income might be included and excluded for each district, making the results not completely comparable.

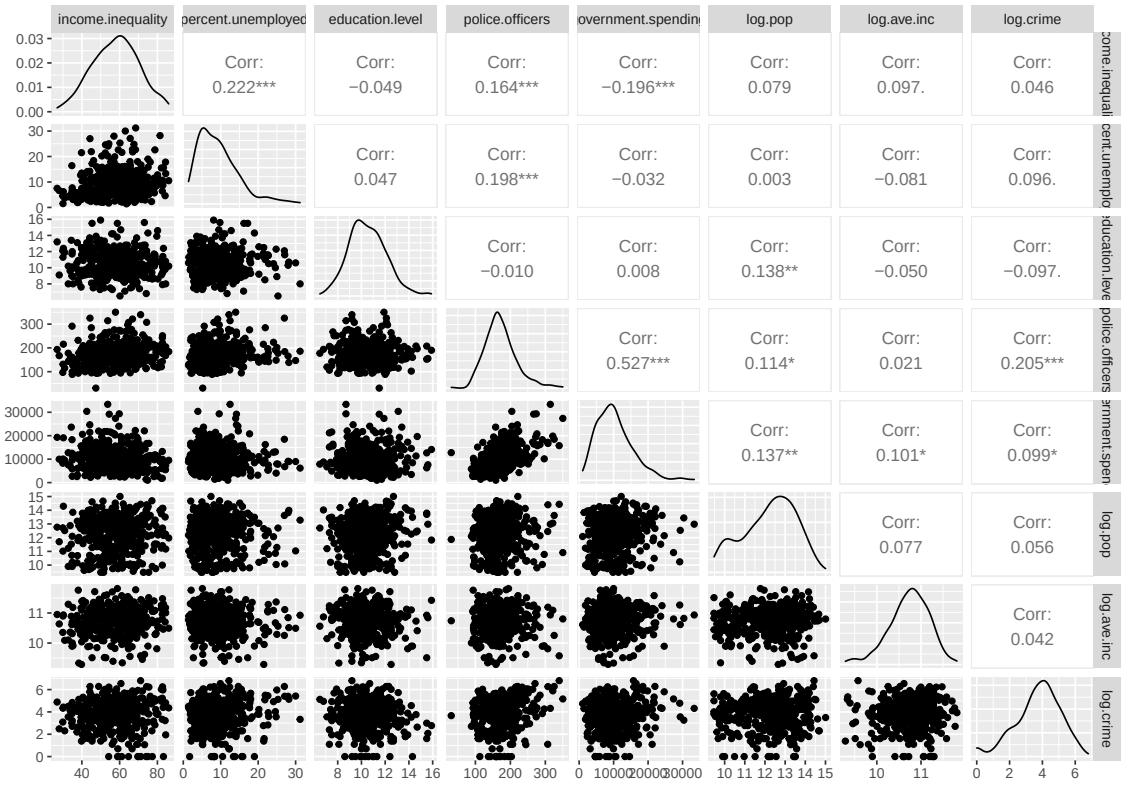
We first look at summary statistics and pairwise scatter plots.

```
   population      average.income   income.inequality   percent.unemployed   education.level   police.officers   government.spending      crime.rate
Min.   :  13000   Min.   :  9700   Min.   :    26.90   Min.   :   -1.700   Min.   :  6.50   Min.   :  31.0   Min.   :      -3298   Min.   : -6.80
1st Qu.:  86000   1st Qu.: 35000   1st Qu.:    49.60   1st Qu.:    5.200   1st Qu.:  9.40   1st Qu.: 142.0   1st Qu.:       6340   1st Qu.: 18.00
Median : 254000   Median : 47700   Median :    58.50   Median :    8.400   Median : 10.40   Median : 165.4   Median :       9652   Median : 47.00
Mean   : 402403   Mean   : 50108   Mean   :    57.91   Mean   :    9.287   Mean   : 10.53   Mean   : 171.9   Mean   :      10264   Mean   : 86.32
3rd Qu.: 542000   3rd Qu.: 63200   3rd Qu.:    66.50   3rd Qu.:   12.200   3rd Qu.: 11.50   3rd Qu.: 194.5   3rd Qu.:      13112   3rd Qu.:106.00
Max.   :3290000   Max.   :135200   Max.   :    86.00   Max.   :   31.200   Max.   : 21.00   Max.   : 350.1   Max.   :      33385   Max.   :897.00
```



We notice several things from the summary statistics and scatterplots. Firstly, there are a number of negative values in some positive variables. These are clearly mistakes, and should be removed. `crime.rate`, `population` and `average.income` clearly have skewed distributions, and would probably benefit from a suitable transformation, such as a log transformation. `percent.unemployed` also has a slightly skewed distribution and might benefit from a transformation, but as a percentage, there is not such a good choice for the transformation, so I will not transform it. There is also a clear outlier in `education.level`, which I will remove. There are six zero values for `crime.rate`, which are not handled by the log-transformation. These could represent values rounded to zero, or might indicate data col-

lection issues. I have removed them for the initial analysis. There is also an outlier in `police.officers`, with the number of police officers much lower than for other districts. I have removed this observation. We now replot the pairwise scatterplots after these transformations.

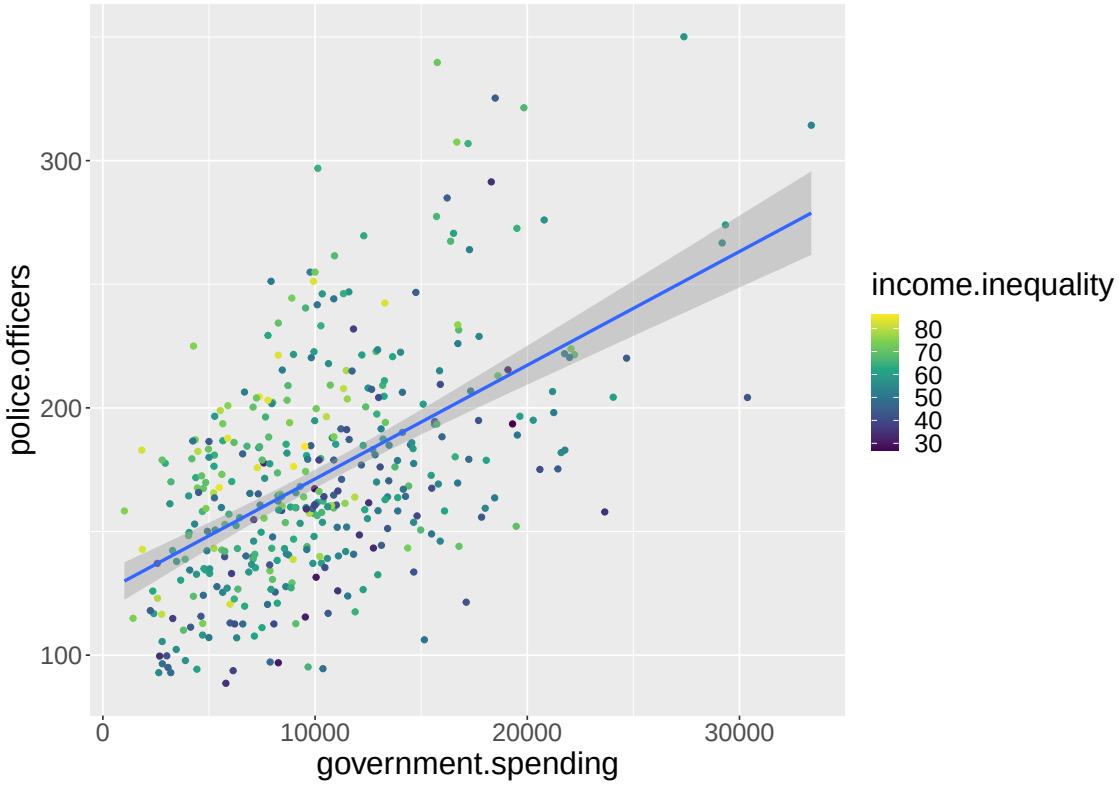


There are also some larger groups of outliers in `government.spending` and `percent.unemployed` that could be removed, but represent enough of the data, and are close enough to the other data points that I prefer to include them.

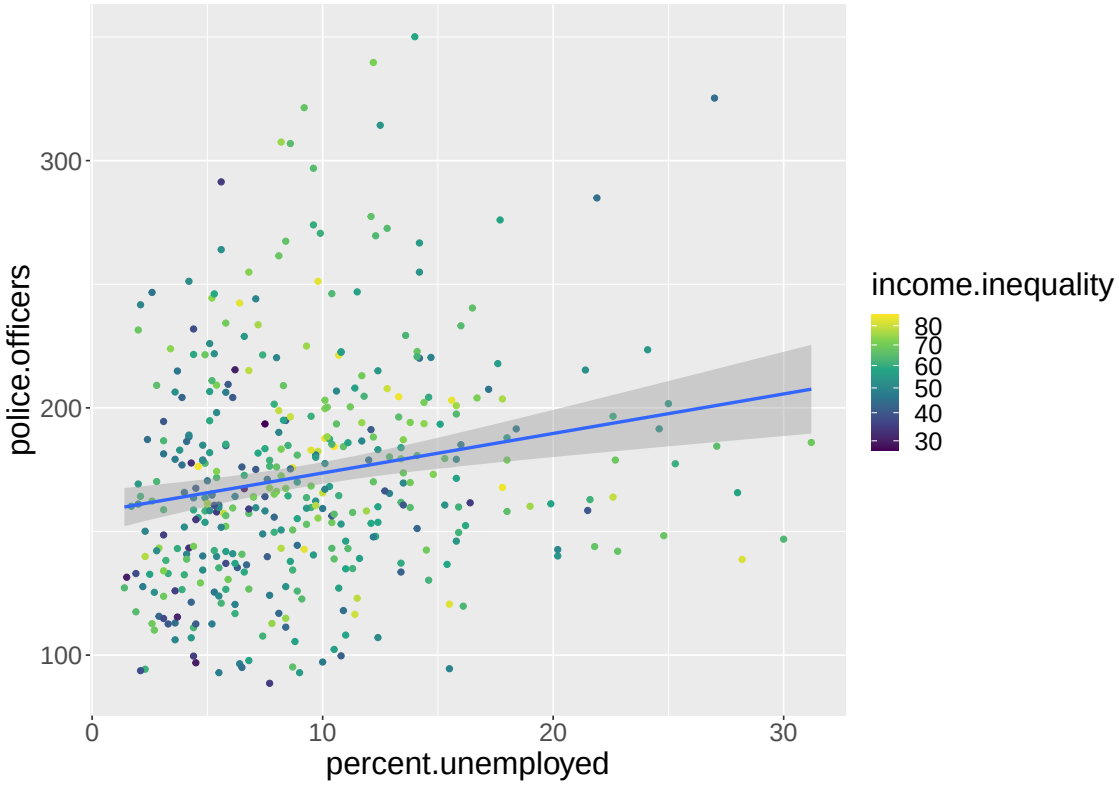
From the pairwise scatterplots, we also note

- There is a fairly strong positive linear relation between `government.spending` and `police.officers`
- There are weak positive associations between `income.inequality`, `unemployment.percent` and `police.officers`.
- There is a weak negative association between `income.inequality` and `government.spending`. This is in spite of both these variables having positive association with `police.officers`.

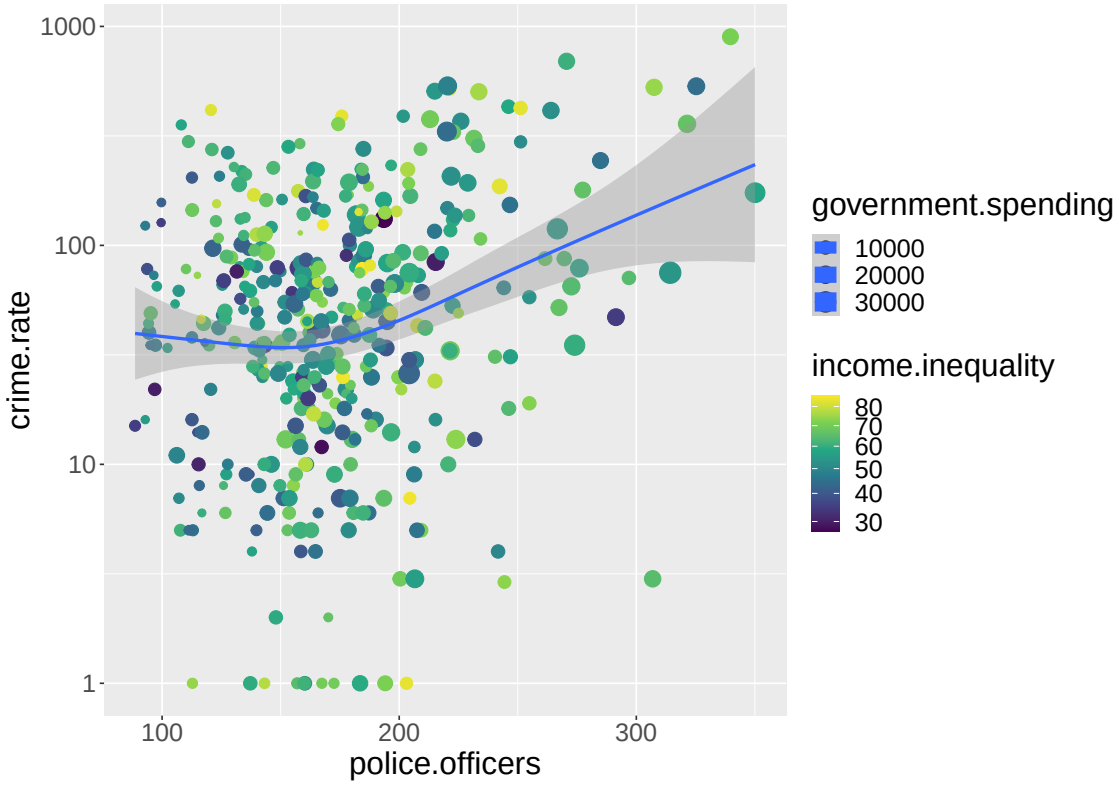
We next look at the relationship between `income.inequality`, `government.spending` and `police.officers`.



and at the relationship between `income.inequality`, `unemployment.percent` and `police.officers`.



Finally, we examine the relation between police officers and crime rate.


1000
100
10
1
crime.rate
100
200
300
police.officers
government.spending
10000
20000
30000
income.inequality
80
70
60
50
40
30

```
HW2Q1<-read.table("HW2Q1.txt")
summary(HW2Q1)
library(GGally)
ggpairs(HW2Q1)
library(dplyr)
ggpairs(HW2Q1%>%filter(education.level<20,
                       percent.unemployed>=0,
                       government.spending>=0,
                       crime.rate>0)%>%
        mutate(log.pop=log(population),
               log.ave.inc=log(average.income),
               log.crime=log(crime.rate))%>%select(-c(population,average.income,crime.rate)))

HW2Q1_good<-HW2Q1%>%filter(
                        education.level<20,
                        percent.unemployed>=0,
                        government.spending>=0,
                        crime.rate>0,
                        police.officers>40)

ggplot(HW2Q1_good,
       mapping=aes(x=government.spending,
                   y=police.officers,
                   colour=income.inequality))+
    geom_point()+
    geom_smooth(method="lm")+
    scale_colour_viridis_c()+
    largertextsize
ggplot(HW2Q1_good,
       mapping=aes(x=percent.unemployed,
                   y=police.officers,
                   colour=income.inequality))+
    geom_point()+
    geom_smooth(method="lm")+
    scale_colour_viridis_c(trans="sqrt")+
    largertextsize

ggplot(HW2Q1_good,
       mapping=aes(x=police.officers,
                   y=crime.rate,
                   colour=income.inequality,
                   size=government.spending))+
    geom_point()+
    geom_smooth(method="gam")+
    scale_colour_viridis_c(trans="sqrt")+
    largertextsize+
    scale_y_log10()
```

Conclusions

- There may be inconsistencies in the data sources. There is also some possibility of bias, though it is not clear in what direction the bias may be.
- There are several outliers that are impossible. These should be removed.
- `population` `average.income` and `crime.rate` are skewed, and may benefit from log transformation.
- There are weak positive associations between `percent.unemployed` `income.inequality` and `police.officers`, and a stronger linear

association between `police.officers` and `government.spending`, and a weak negative association between textttincome.inequality and `government.spending`.

- There is a non-linear relation between `police.officers` and log(`crime.rate`).

2. *The file `HW2Q2.txt` contains the following data from a university's international development office about trends in overseas student applications*

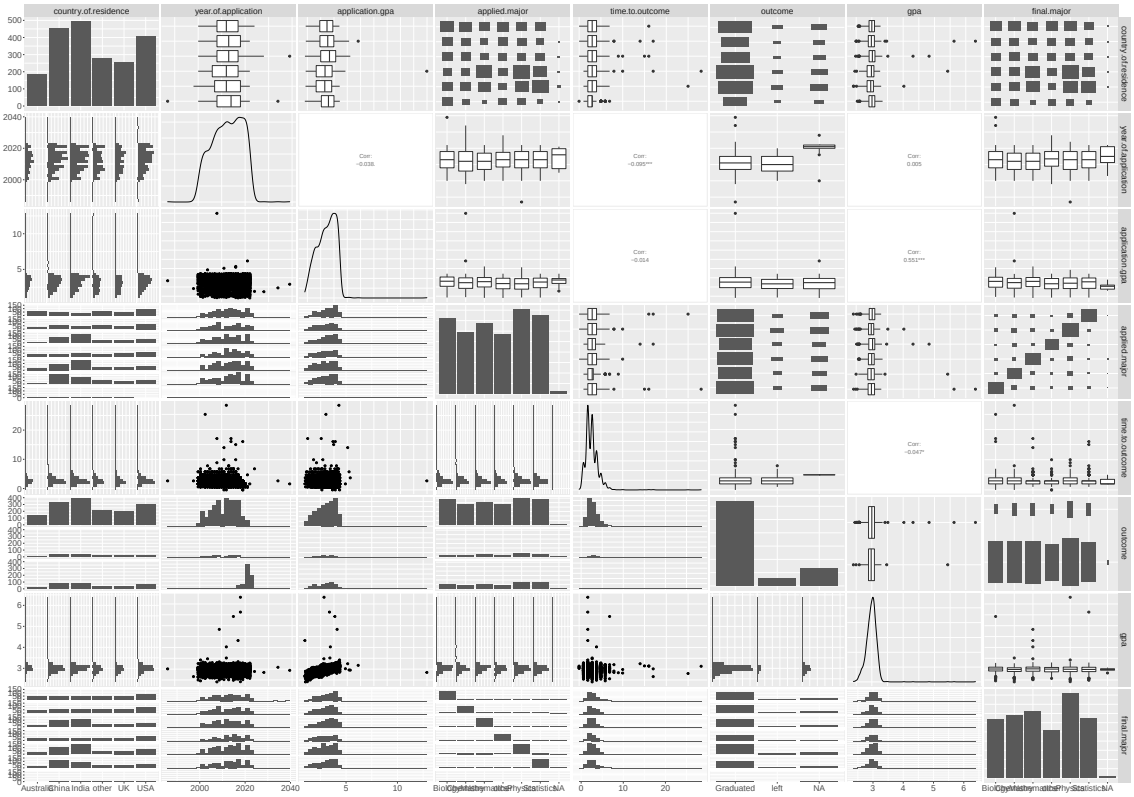
Variable	*Meaning*
country.of.residence	*The country of the applicant's residence*
year.of.application	*The year the application was made*
application.gpa	*The GPA at time of application*
applied.major	*The major to which the student applied*
time.to.outcome	*The time the student spent at the university*
outcome	*The result of the student's studies*
gpa	*The student's final GPA*
final.major	*The students declared major at the time they left the university*

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

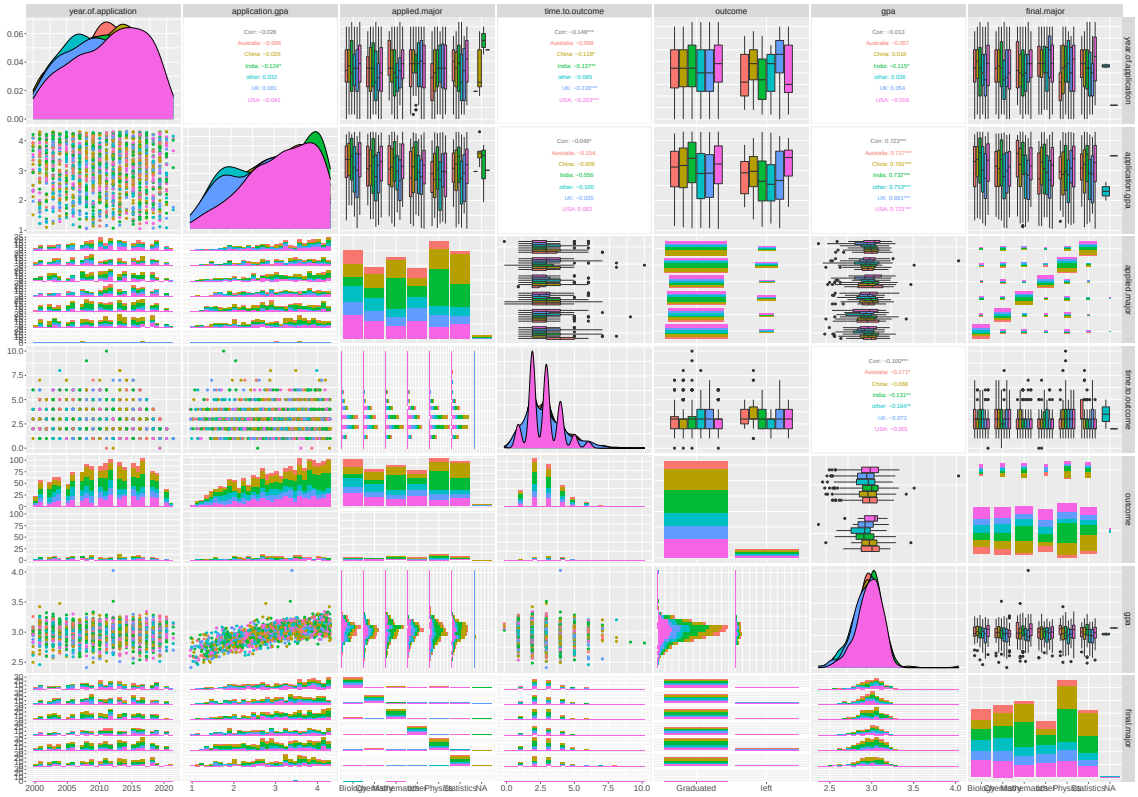
We first consider the source of the data. Since it should come from the university's own records, the data about the student's performance at the university should be unbiased and reliable. The data obtained from the application — `country.of.residence` and `application.gpa` could be false. We would expect the university to require some proof of these, and so they are unlikely to be very biased. There may be different standards for different applicants, so the application GPAs may not all be comparable. There may also be selection bias, as individuals will choose whether to apply based on their GPA. However, as the population of interest is students who apply to the university, this may not be too serious.

We start with a summary of the data, and pairwise scatterplots.

```
country.of.residence   year.of.application   application.gpa    applied.major    time.to.outcome      outcome          gpa          final.major
Australia:       182   Min.   :       1987   Min.   : 1.008   Biology    :360   Min.   :  0.000   Graduated:1594   Min.   :2.411   Biology    :319
China    :       453   1st Qu.:       2008   1st Qu.: 2.381   Chemistry  :292   1st Qu.:  2.000   left     : 138   1st Qu.:2.893   Chemistry  :340
India    :       494   Median :       2013   Median : 3.122   Mathematics:337   Median :  3.000   NA's     : 331   Median :2.997   Mathematics:360
other    :       276   Mean   :       2012   Mean   : 3.029   other      :283   Mean   :  2.942                      Mean   :2.986   other      :256
UK       :       253   3rd Qu.:       2018   3rd Qu.: 3.745   Physics    :404   3rd Qu.:  4.000                      3rd Qu.:3.083   Physics    :460
USA      :       405   Max.   :       2039   Max.   : 12.485  Statistics :373   Max.   : 28.000                      Max.   :6.310   Statistics :322
                                                               NA's       : 14   NA's   :    330                      NA's   :   6
```



We immidiately see that there are a number of `NA` values. Many of these are for the variables `outcome` and `time.to.outcome`, which may be for students who are still enrolled. Indeed the vast majority of these individuals are recent applicants, which supports this explanation. There are also a number of outliers in some variables. In some cases `year.of.application` is in the future, which suggests a mistake. Given the majority of the data are from 2000 onwards, earlier application dates are probably also a mistake. There are also impossible values for `application.gpa`, probably reflecting institutions with a different system for calculating GPA. These should be removed. There are also large outliers in `time.to.outcome`, which may be genuine exceptional cases, or may be mistakes. There are some cases where `application.year` + `time.to.outcome` is in the future. These are clearly mistakes, and should be removed. After removing these cases, we replot the scatterplots. We also use colour to indicate the `country.of.residence`.

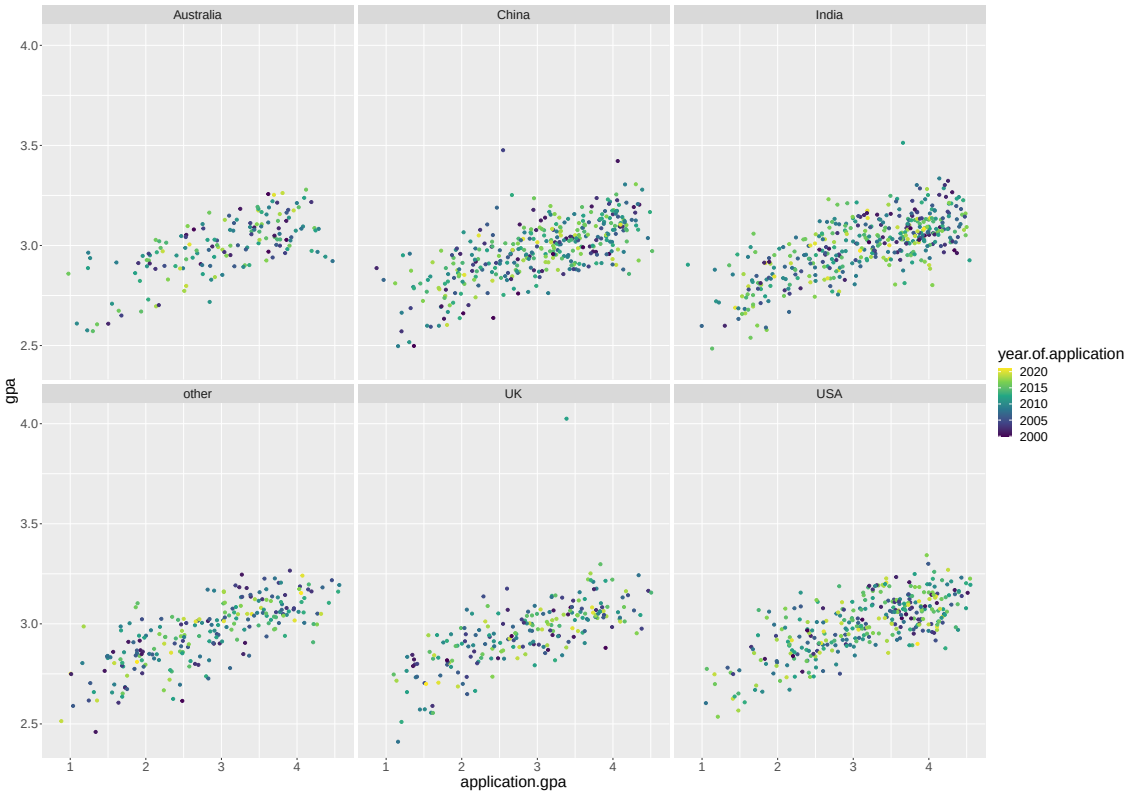


From the new scatterplots, we see that there is some correlation between `year.of.application` and `time.to.outcome`. This may be caused by sampling bias — more recent applicants with long time to outcome are still enrolled, and so `time.to.outcome` will be `NA`. This will cause spurious correlation between these variables. If we restrict to students with `year.of.application` before 2012, the correlation is reduced to -0.02521638. There is also strong linear correlation between `application.gpa` and `gpa`. There are several clear outliers in `gpa`, which we might consider removing if they significantly alter the results of the analysis. We also see that in 74.0% of cases where both are known, `application.major` and `final.major` are the same. There are therefore a very limited number of cases for any particular change of major, so we are unlikely to be able to make any firm conclusions about this. We might consider removing one of these variables, and perhaps introducing a new boolean variable indicating whether the student changed their major subject. If using a method for variable selection, we might leave both variables in the dataset, and allow the variable selection method to determine which is the better predictor.

We see that `country.of.residence` is associated with `application.major`, `application.gpa` and `final.gpa`. The `outcome` variable is very unbalanced, with most students graduating. Only 137 students have outcome

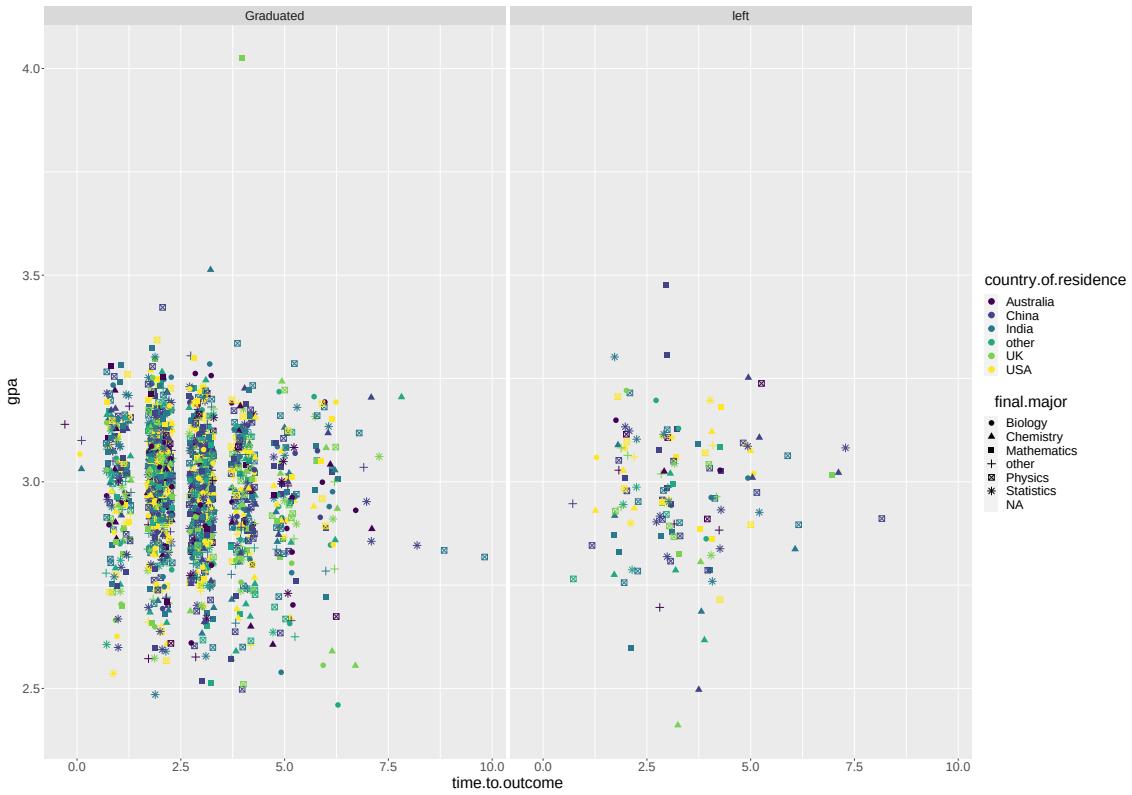
"left". This limits our ability to detect patterns in these students. There is also a negative relation between `time.to.outcome` and `gpa`, which makes sense. This seems to vary by `country.of.residence`.

Looking in more detail at the relation between `application.gpa` and `gpa`,



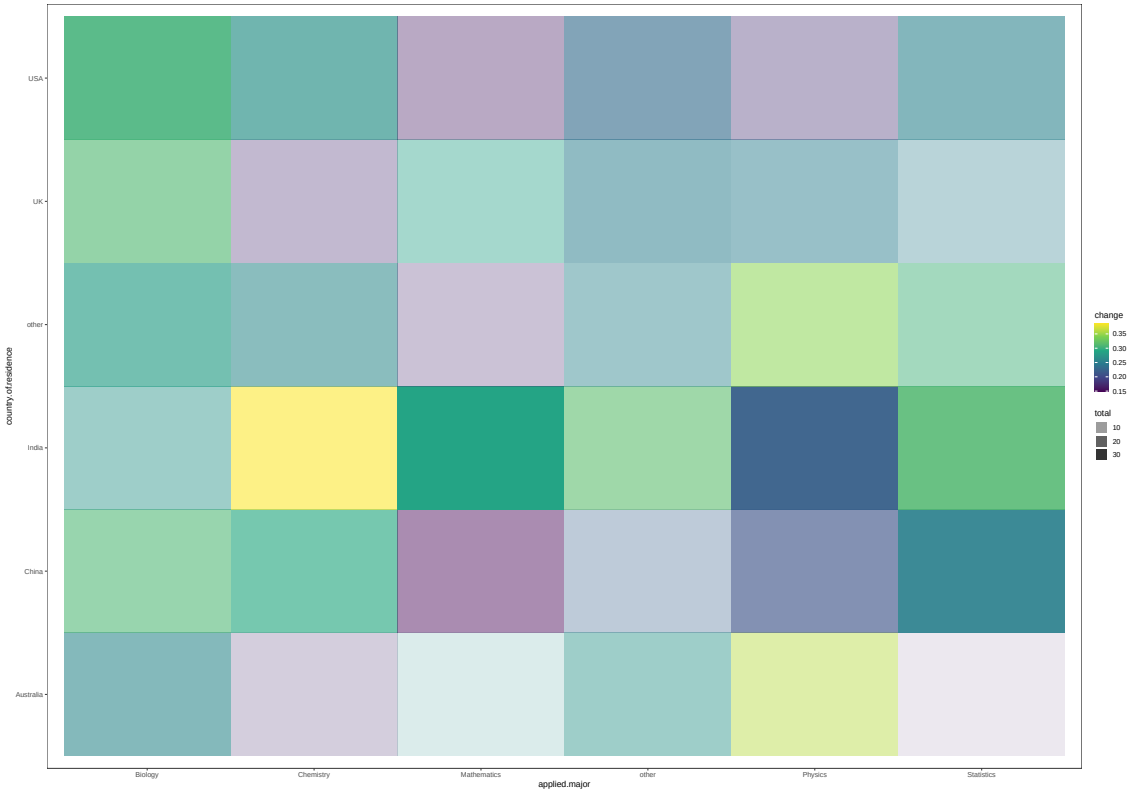
we see there is a strong linear relation with slope approximately 0.12, which does not seem to be affected by other factors. This suggests we could look at $\texttt{gpa} - 0.12\texttt{application.gpa}$ as a new predictor to indicate the students performance at the university, relative to initial expectations.

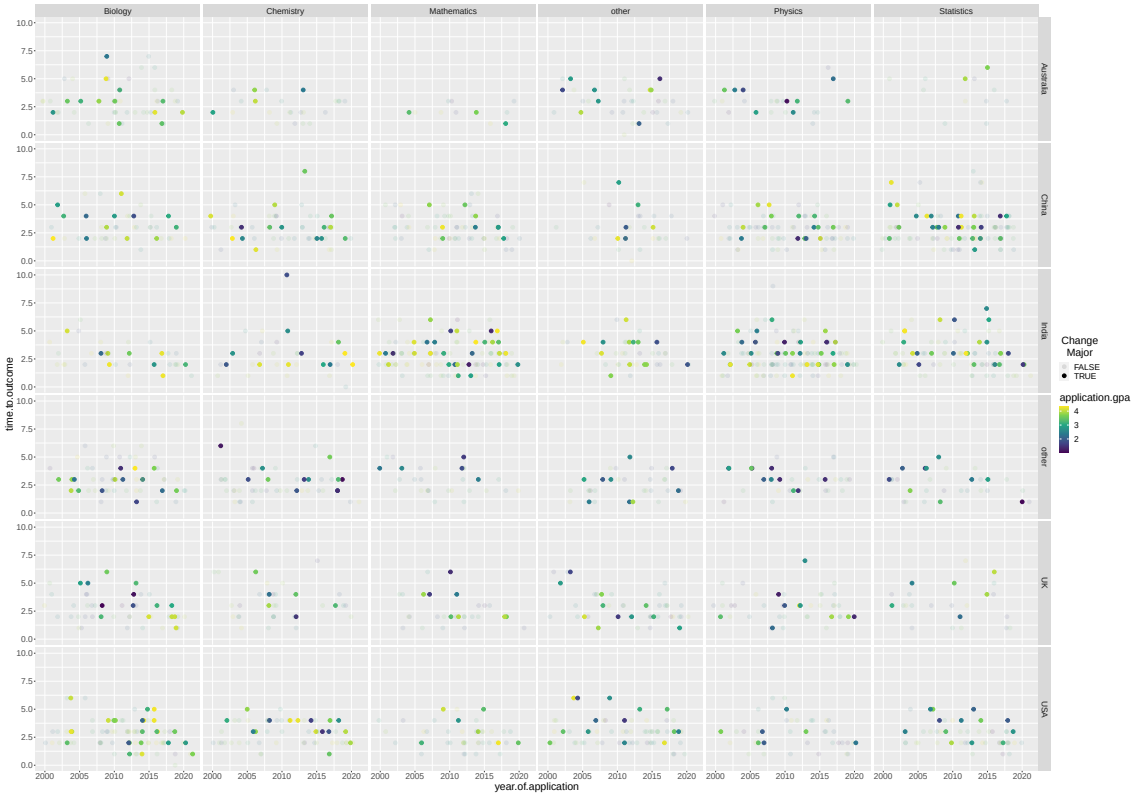
We look to see whether there are any patterns among students who leave:



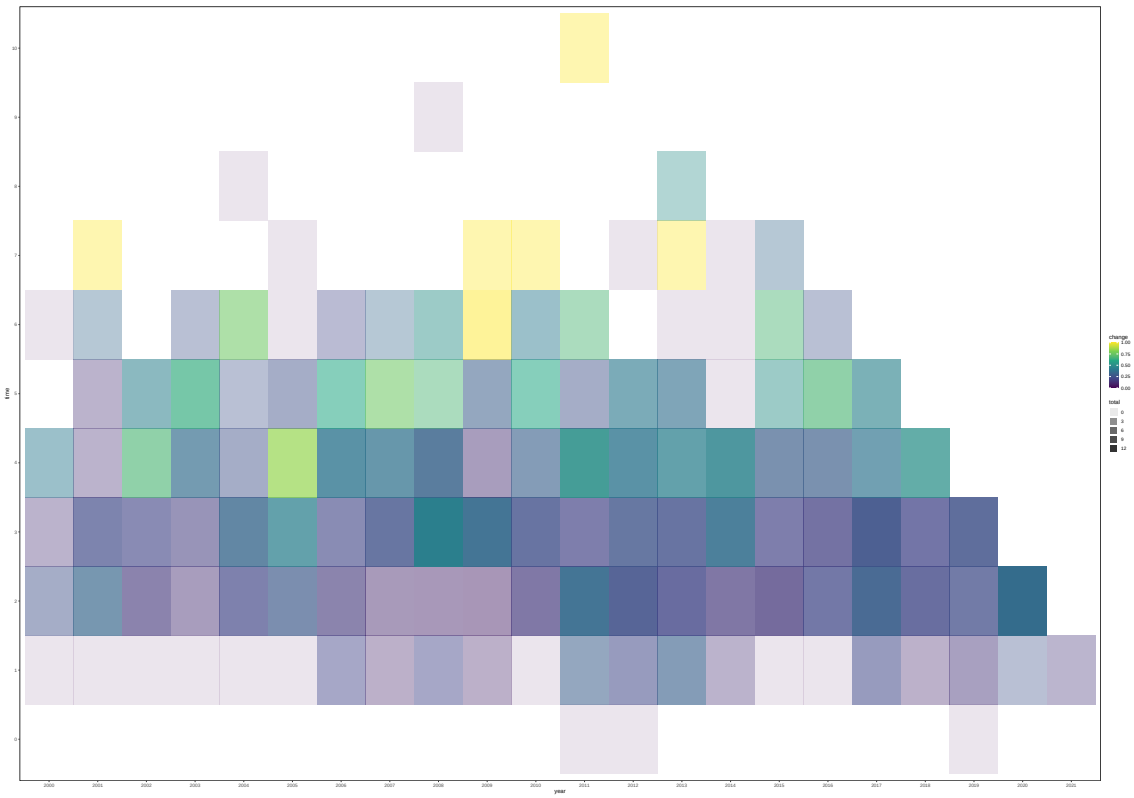
This plot does not show any clear patterns. [An alternative approach would be to plot both students who leave and students who graduate on the same plot, using alpha or size, or some other method to distinguish them.]

We can also look at the proportion of students who change major.


USA
UK
other
India
China
Australia
country.of.residence
Biology
Chemistry
Mathematics
other
Physics
Statistics
applied.major
change
0.35
0.30
0.25
0.20
0.15
total
10
20
30



The first plot gives a summary based on the categorical variables. The second plot gives quite a bit of additional information, but at the cost of making it harder to judge the overall effect of the categorical variables. From these plots, we see that the probability of changing major varies in a complicated way with `country.of.residence` and `applied.major`. The second plot shows that the probability of changing major seems to be increase with `application.gpa` and with `year.of.application`. This trend seems to be the same for all combinations of `country.of.residence` and `applied.major`. We look in more detail at the relation between changing major and `year.of.application` and `time to outcome`.



We see that the probability of changing major seems to increase with `time to outcome`, which makes intuitive sense. There does not appear to be a strong relation between probability of changing major and `year.of.application`.

Conclusions

- There are a number of NA values. Some represent students in progress, and should possibly be changed to indicate this.
- There are some clear data issues, such as application dates in or time to outcome in the future, or GPA outside the range for GPA.
- There are also some outliers in `year.of.application` and `time.to.outcome`, which are likely to be mistakes.
- There are also several outliers in `gpa`. It is unclear whether these need to be removed for the analysis.
- The data are censored — `time.to.outcome` is only available in cases where the degree is completed. This needs to be considered in analysing the data, as it generates spurious correlation between `year.of.application` and `time.to.outcome`.
- There is strong linear correlation between `application.gpa` and `gpa`.

- `application.major` and `final.major` are the same in a large majority of cases. We may consider adding a feature `change.major` to indicate whether a student changes major.
- A large majority of completed students graduate. It is hard to detect any patterns in which students leave.
- `country.of.residence` is associated with `application.gpa`, `applied.major`, `final.major` and `final.gpa`
- There is a weak negative relation between `time.to.outcome` and `gpa`.
- The probability of a student changing majors varies with `country.of.residence` and `applied.major`. It also increases with `gpa` and `time.to.outcome`.

```
HW2Q2<-read.table("HW2Q2.txt")
HW2Q2$country.of.residence<-as.factor(HW2Q2$country.of.residence)
HW2Q2$applied.major<-as.factor(HW2Q2$applied.major)
HW2Q2$outcome<-as.factor(HW2Q2$outcome)
HW2Q2$final.major<-as.factor(HW2Q2$final.major)
options(width=200)
summary(HW2Q2)
library(GGally)
ggpairs(HW2Q2)+largertextsize
table(HW2Q2$year.of.application[is.na(HW2Q2$time.to.outcome)])
library(dplyr)
HW2Q2_good<-HW2Q2%>%filter(year.of.application>=2000&
                           year.of.application<=2022&
                           application.gpa>=0&
                           gpa>=0&
                           application.gpa<=4.3&
                           gpa<=4.3&
                           year.of.application+time.to.outcome<=2022)
ggpairs(HW2Q2_good%>%select(-c("country.of.residence")),
        mapping=aes(colour=HW2Q2_good$country.of.residence))+
    largertextsize
ggplot(HW2Q2_good,mapping=aes(x=application.gpa,
                              y=gpa,colour=year.of.application))+
    geom_jitter(width=0.3,height=0)+
    facet_wrap(country.of.residence~.)+
    scale_colour_viridis_c()+
    largertextsize
ggplot(HW2Q2_good,mapping=aes(x=time.to.outcome,
                              y=gpa,colour=country.of.residence,
                              shape=final.major))+
    geom_jitter(width=0.3,height=0,size=3)+
    facet_wrap(outcome~.)+
    scale_colour_viridis_d()+
    largertextsize

ggplot(HW2Q2_good%>%filter(!is.na(applied.major)&!is.na(final.major))%>%
       group_by(country.of.residence,applied.major)%>%
       summarise(total=sum(final.major!=applied.major),
### Make tiles with fewer points more transparent as their results are
### less reliable. We could use total=n() to use the total number of
### points in the tile, but as major changes are fairly rare, the
### variability of the estimated probability depends more on the
### number of changes. This approach needs a colour scale that depends
### on hue, rather than light/dark, as light/dark will conflict with
### alpha.  Even in this case, it is not very easy to judge the
### significance of each tile.
                 change=mean(final.major!=applied.major)),
       mapping=aes(x=applied.major,
                   y=country.of.residence,
                   fill=change,
                   alpha=total))+
    geom_tile()+
    scale_fill_viridis_c()+
    largertextsize+
    scale_alpha_continuous(trans="sqrt")+
    theme_bw()+ ### remove background colour and gridlines to make tile
    theme(panel.grid.major=element_blank()) ### colour and alpha easier to see.

ggplot(HW2Q2_good%>%filter(!is.na(applied.major)&!is.na(final.major)),
       mapping=aes(x=year.of.application,
                   y=time.to.outcome,
                   alpha=final.major!=applied.major,
                   colour=application.gpa))+
    geom_jitter(width=0.3,height=0,size=3)+
    scale_colour_viridis_c()+
    largertextsize+
    facet_grid(country.of.residence~applied.major)+
    scale_alpha_discrete(name="Change\nMajor")

ggplot(HW2Q2_good%>%filter(!is.na(applied.major)&!is.na(final.major))%>%
       mutate(year=as.factor(year.of.application),# convert to factors
              time=as.factor(time.to.outcome))%>% # in order to group.
       group_by(year,time)%>%
       summarise(total=sum(final.major!=applied.major),
                 change=mean(final.major!=applied.major)),
       mapping=aes(x=year,
                   y=time,
                   fill=change,
                   alpha=total))+
    geom_tile()+
    scale_fill_viridis_c()+
    largertextsize+
```

3. *The file* `HW2Q3.txt` *contains the following data from an electricity company, who are trying to forecast electricity demand in each city.*

Variable	*Meaning*
city	*The city being supplied.*
year	*The year.*
population	*The population at the time.*
average.temp	*The average daytime high temperature over the year.*
rainfall	*The total annual rainfall.*
price.adj	*CPI-adjusted price per KWh of electricity*
consumption	*Total electricity consumption*

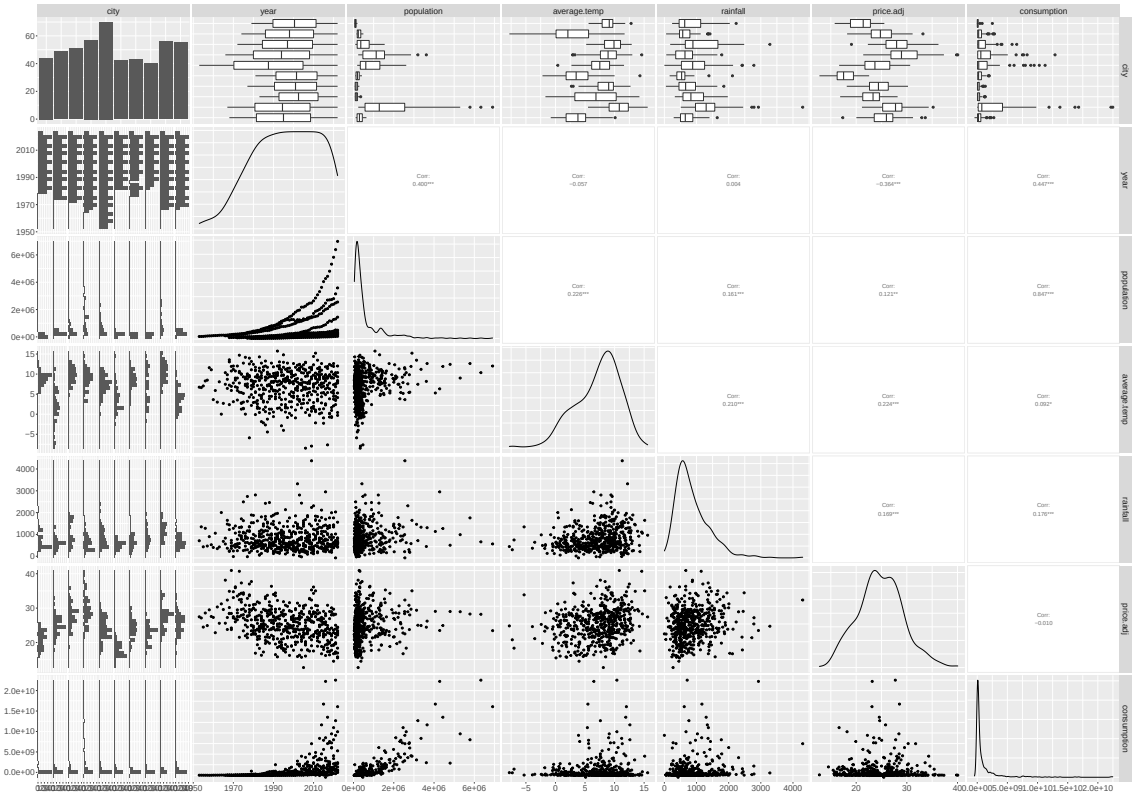
Electricity consumption is from the company's meters. Population is from government census data. Weather data is from government historical weather records. Pricing data is from the company's records, and the CPI adjustment is using government economic data.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

The data are collected from a range of different sources, so there is always the possibility of inconsistencies in the data. For example, are the city boundaries used in government census data constant, and do they correspond to the areas considered part of the city by the electricity company. The sources used should all be reasonably reliable, with limited reason to suspect bias.

We start with a summary of the data and pairwise scatterplots.

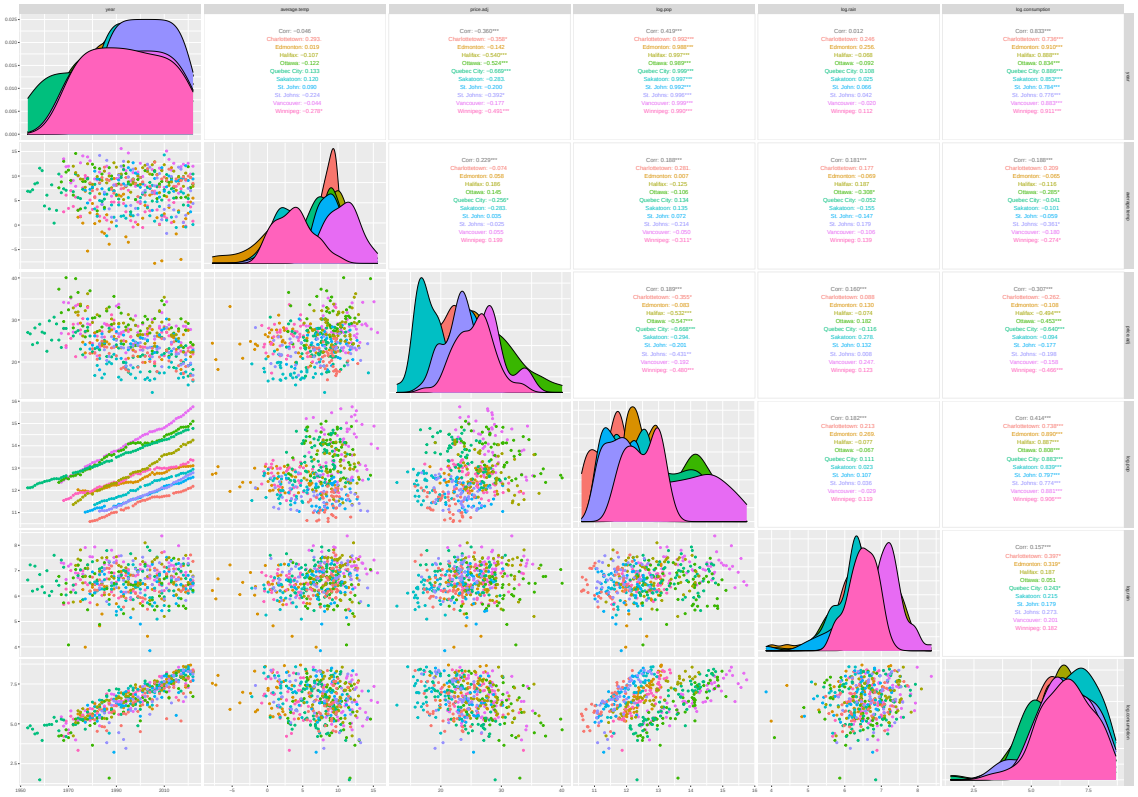
```
     city           year         population       average.temp       rainfall        price.adj         consumption
Quebec City: 70   Min.   :1953   Min.   :  40000   Min.   :-7.800   Min.   :   1.0   Min.   :12.75   Min.   :-4.600e+08
Ottawa     : 57   1st Qu.:1984   1st Qu.: 152000   1st Qu.: 4.650   1st Qu.: 501.0   1st Qu.:21.98   1st Qu.: 5.000e+07
Vancouver  : 56   Median :1997   Median : 288000   Median : 7.800   Median : 738.0   Median :24.58   Median : 2.050e+08
Winnipeg   : 55   Mean   :1996   Mean   : 628649   Mean   : 7.109   Mean   : 869.6   Mean   :24.83   Mean   : 1.090e+09
Halifax    : 51   3rd Qu.:2010   3rd Qu.: 642000   3rd Qu.: 9.800   3rd Qu.:1126.0   3rd Qu.:27.70   3rd Qu.: 9.080e+08
Edmonton   : 49   Max.   :2022   Max.   :6911000   Max.   :15.600   Max.   :4303.0   Max.   :40.05   Max.   : 2.248e+10
(Other)    :169                                                     NA's   :     2
```



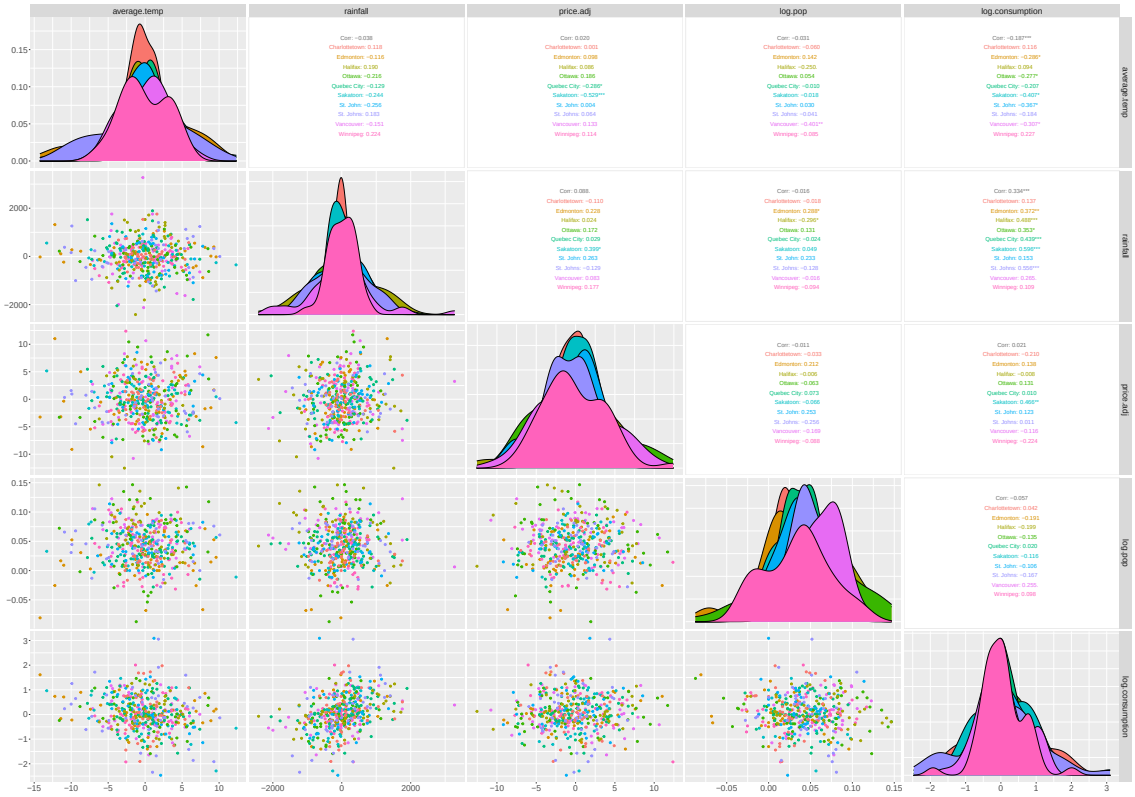
From this, we notice several things. Firstly, there are some negative values of consumption, which are clearly wrong. There are 2 NA values for `price.adj`. We also note that population, rainfall and consumption have skewed long-tailed distributions. Population and consumption are highly correlated, which is to be expected, and suggests creating the feature $\texttt{per.capita.consumption} = \frac{\texttt{consumption}}{\texttt{population}}$. We also note that different cities have very different values of the predictors. Population appears to grow exponentially over time for each city, and the relation appears to be very strong for each city.

We remove the negative consumption values and the NA values in `price.adj`, replace consumption by per.capita.consumption, and log-transform population and rainfall. We then redraw the pairwise scatterplots coloured by city.

This plot immediately shows a huge outlier in per-capita consumption, which need to be removed. If we look at per-capita consumption against year, there is a second large outlier that also needs to be removed. There is a big outlier in rainfall on the log-transformed scale, where the annual rainfall was only 1mm. This is either an extremely severe drought, or an error in the data. In either case, we remove this point from the data. After replotting the pairwise scatterplots, we see that the second highest per-capita consumption is also an outlier for its city and year, and that per-capita consumption is still skewed and exponentially increasing over time, suggesting a log-transformation. Removing this outlier and log-transforming the per-capita consumption gives the following pairwise scatterplots.



These plots make clear that log(*population*) and log(`per.capita.consumption`) are both increasing approximately linearly over time. (Equivalently, population and per-capita consumption are both growing exponentially). We also see a decreasing trend in adjusted price over time. Because of the strong correlations with time, it is hard to judge the relations with other variables. One approach for removing these trends is to take the difference between consecutive years. In this dataset there are some missing years, which would cause issues for using this technique in an analysis, but shouldn't prevent its use for data exploration.



From this plot, we see that the per-capita consumption grows at approximately the same rate in all cities, and that this rate is weakly positively correlated with changes in rainfall, and weakly negatively correlated with changes in average temperature. There is no strong relation between the change in adjusted price and relative change in per-capita consumption, or between relative change in population and relative change in per-capita consumption.

Conclusions

- There are several invalid and missing values that should be removed.
- There are several outliers that should be removed or separately analysed.
- Population and consumption are highly correlated. A natural solution is to calculate per-capita consumption.
- Per-capita consumption and population are both growing exponentially over time in each city.
- Per-capita consumption, population and rainfall have skewed distributions in each city, and a log transformation may be appropriate.

- The relative growth in per-capita consumption is weakly positively correlated with rainfall, and weakly negativel correlated with average temperature.

```
HW2Q3$city<-as.factor(HW2Q3$city)
summary(HW2Q3)

ggpairs(HW2Q3)+largertextsize

HW2Q3_valid<-HW2Q3%>%filter(consumption>0&!is.na(price.adj))%>%
    mutate(per.capita.consumption=consumption/population)%>%
    select(-c("consumption"))

ggpairs(HW2Q3_valid%>%mutate(log.pop=log(population))%>%
        select(-c("city","population")),mapping=aes(colour=HW2Q3_valid$city))

HW2Q3_good<-HW2Q3_valid%>%filter(per.capita.consumption<6000&
                                 rainfall>5&
                                 (per.capita.consumption/(log(population)-10))<3000)

ggpairs(HW2Q3_good%>%mutate(log.pop=log(population),log.rain=log(rainfall),log.consumption=log(per.capita.consumption))%
        select(-c("city","population","rainfall","per.capita.consumption")),mapping=aes(colour=HW2Q3_good$city))

HW2Q3_vgood<-HW2Q3_valid%>%filter(per.capita.consumption>10&
                                  per.capita.consumption<6000&
                                  rainfall>5&
                                  (per.capita.consumption/(log(population)-10))<3000)

HW2Q3_trans<-HW2Q3_vgood%>%mutate(log.pop=log(population),log.consumption=log(per.capita.consumption))%>%
    select(-c("population","per.capita.consumption"))

HW2Q3_diff<-HW2Q3_trans[-1,]-HW2Q3_trans[-489,]

HW2Q3_diff$city<-HW2Q3_trans$city[-1]

HW2Q3_diff<-HW2Q3_diff%>%filter(year==1)

HW2Q3_diff<-HW2Q3_diff%>%select(-c("year"))


ggpairs(HW2Q3_diff%>%select(-c("city")),mapping=aes(colour=HW2Q3_diff$city))+largertextsize
```

4. *An pensions company is modelling improvements in mortality. It collects the following data on its policyholders:*

Variable	*meaning*
init.age	*The age of the policyholder at the time of plan initiation.*
init.year	*The year of plan initiation*
death.age	*The age of the policyholder at death (0 if policyholder is still alive)*
death.year	*The year of the policyholder's death*
sex	*The sex of the policyholder*
race	*The race of the policyholder*
income	*The policyholder's income (adjusted for inflation) at time of initiation.*
smoking	*Whether and how much the policyholder smokes at time of initiation.*
health	*A measure of the policyholder's overall health at time of initiation, with 100 representing perfect health and 0 being*

The data are in the file `HW2Q4.txt`.

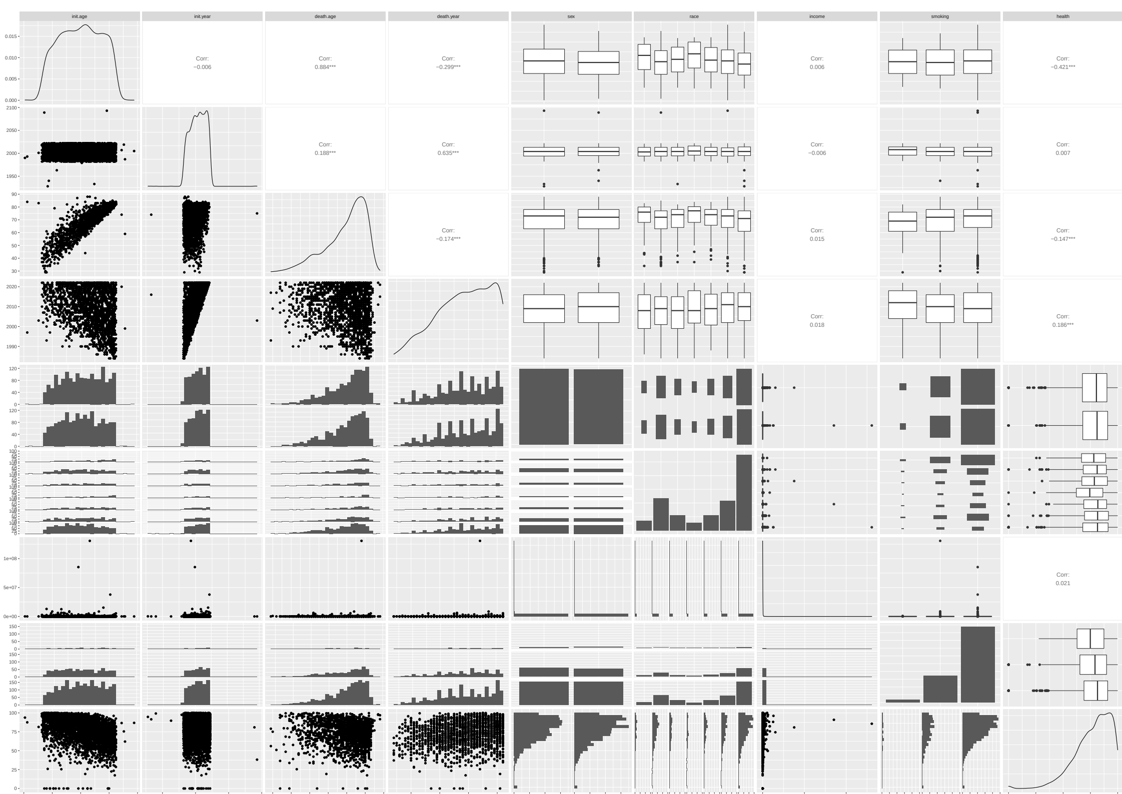
The data are from the pension company's records.

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.

Data from the company's records should be unbiased, and for setting premiums and payments, the company will do a lot to ensure accuracy. There is sampling bias relative to the population, as this consists of individuals with pensions. However, this is the population of interest. There could be issues if there are changes to which individuals have a pension over time. Some of the data are redundant, since initial age, age at death, initial year and year of death should follow a linear relation (up to rounding). This redundancy can be used as a check to remove errors in the data. The use of zero to represent NA is bad and needs to be corrected.

We begin with a summary of the data and pairwise scatterplots.

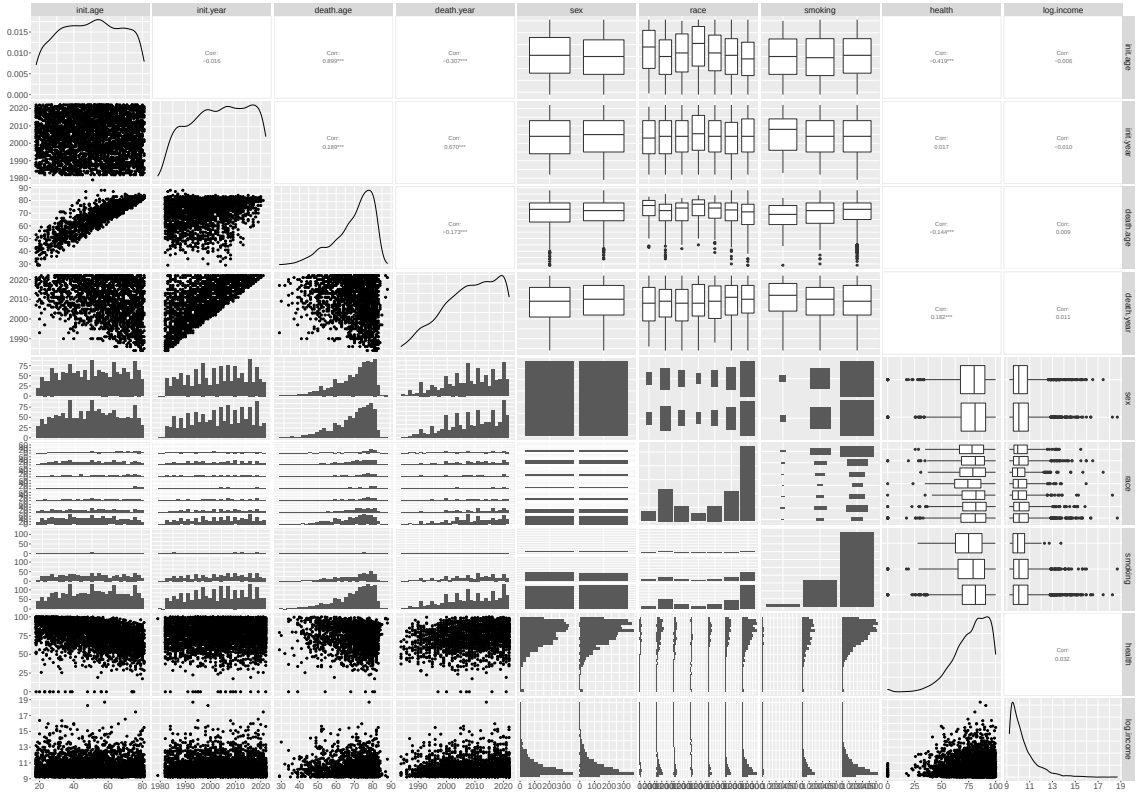
```
   init.age       init.year       death.age       death.year          sex               race                  income              smoking          health
 Min.   : 1.0   Min.   :1928   Min.   :29.00   Min.   :1984   Female:      1699   Black      :   175   Min.     :     10000   heavy:      82   Min.   : 0.00
 1st Qu.:35.0   1st Qu.:1994   1st Qu.:63.00   1st Qu.:2002   Male  :      1683   East Asian:   589   1st Qu.  :     14200   light:     863   1st Qu.:68.00
 Median :50.0   Median :2004   Median :72.00   Median :2009         :   Hispanic   275        : Median   22900    :Non-smoker   2437 : Median    80.80  :
 Mean   :49.8   Mean   :2004   Mean   :69.43   Mean   :2009         : Indiginous   137        :   Mean   198459   :                      Mean    77.62  :
 3rd Qu.:65.0   3rd Qu.:2013   3rd Qu.:78.00   3rd Qu.:2017         :      other   282        :3rd Qu.   50800    :                   3rd Qu.    90.40  :
 Max.   :97.0   Max.   :2093   Max.   :88.00   Max.   :2022         :South Asian   545        :   Max.   130741500:                      Max.    100.00 :
                               NA's   : 1601   NA's   :1601                        White      :  1379
```



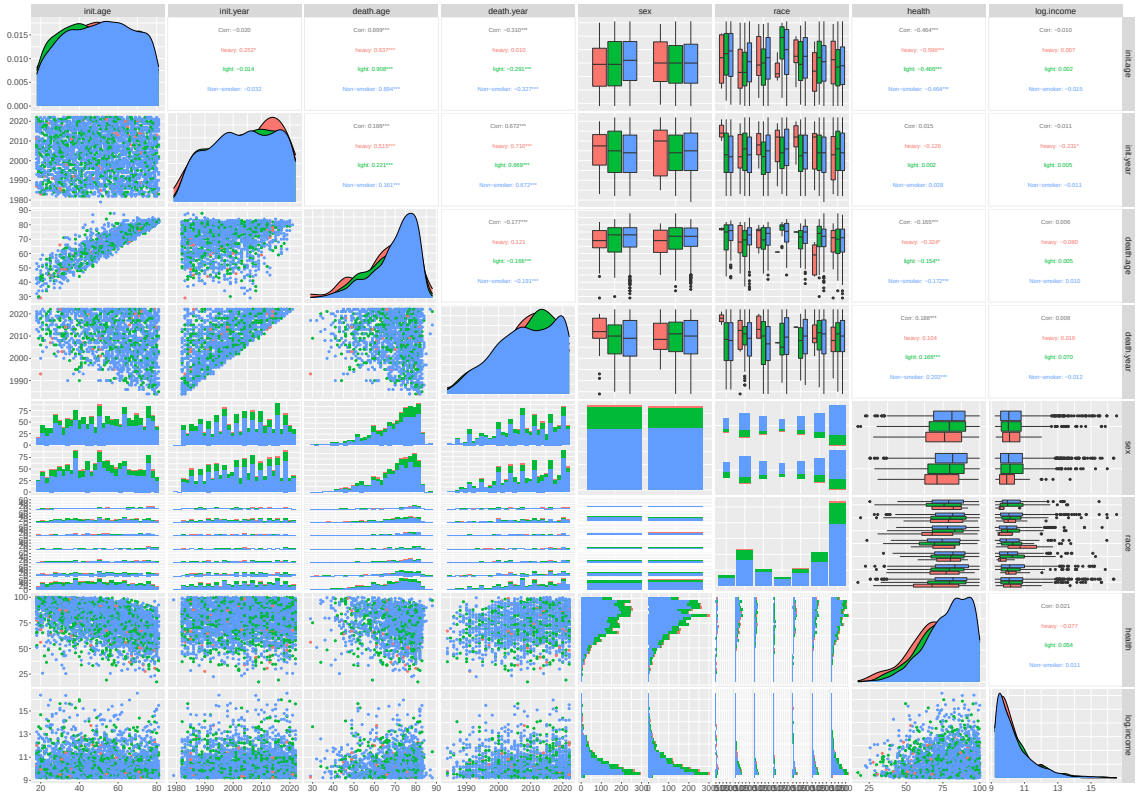
We see that there are several outliers for each variable. Some are clearly impossible, such as initial years in the future, or death age less than initial age, while others are possible, but seem unlikely, such as initial ages less than 18. We see that `income` is very positively skewed, while `death.age`, `death.year` and `health` are negatively skewed. It would clearly be appropriate to log transform `income`. There are also several outliers in `income`, but these may not be outliers after transformation. We therefore start by removing all data points with `init.age` < 18, `init.age` > 85, `init.year` < 1970, `init.year` > 2022, or

$$|\texttt{death.age} - \texttt{init.age} - (\texttt{death.year} - \texttt{init.year})| > 1$$

(inconsistent, even with rounding).

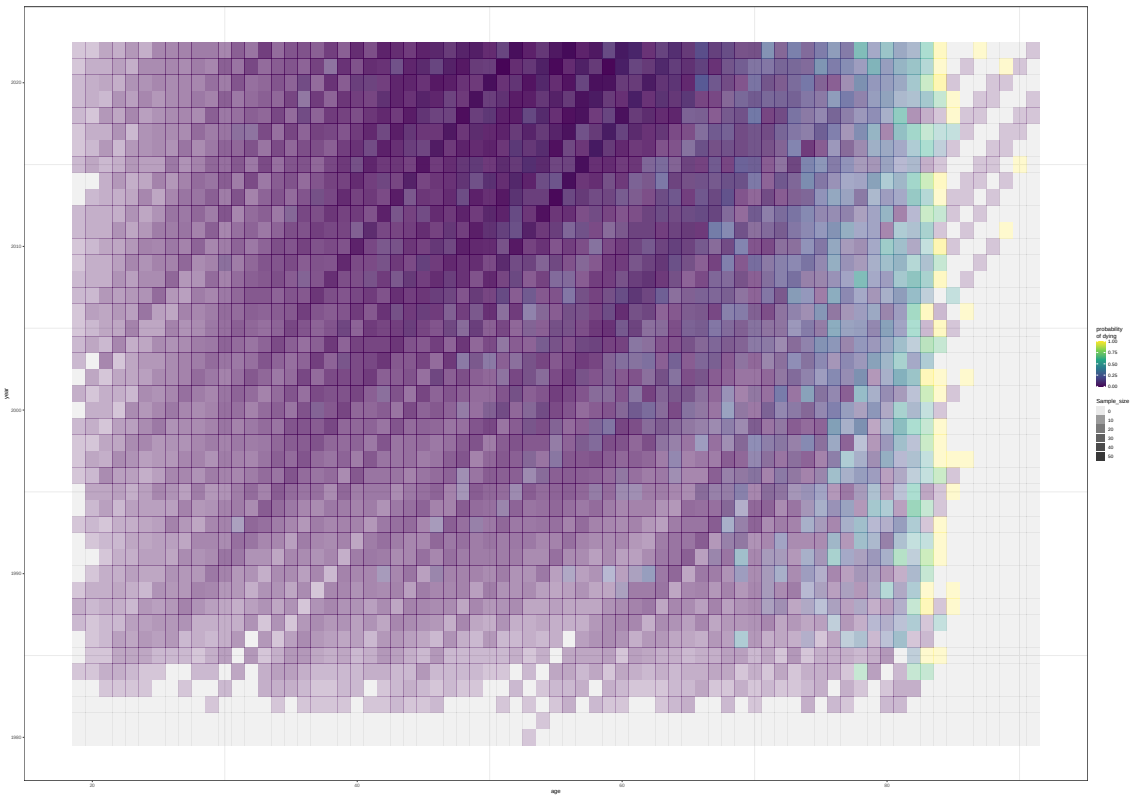


Even after log-transformation, the largest 3 incomes are clear outliers, and there are several 0 values for health, which are also outliers. We remove these and redraw the pairwise scatterplots, using colour to represent smoking status.

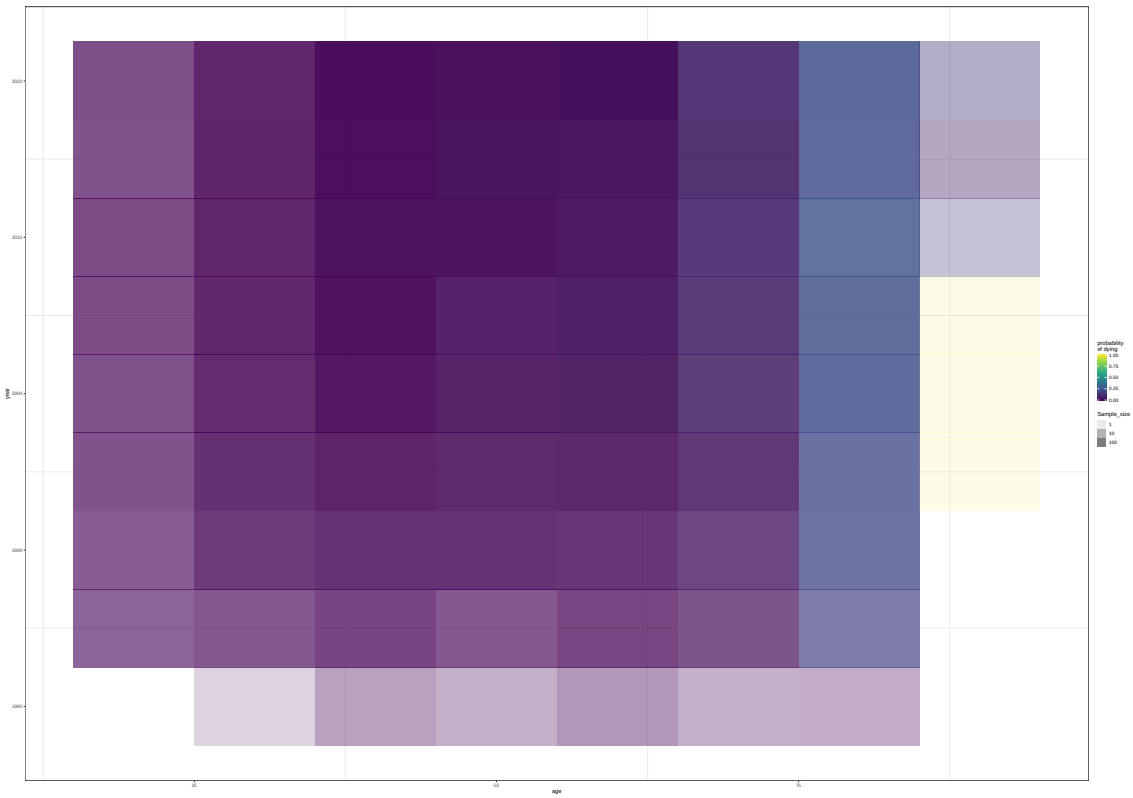


We see there is high correlation between `init.age` and `death.age`. This is partially due to sampling bias, because `death.age` must be more than `init.age`, and also because of censorship. The earliest pensions started in 1980, so if the beneficiaries were 20 at the time, then they would be only 63 now, so could not have a `death.age` more than 63. `death.age` is also correlated with `init.year` and `death.year`, and these are also impacted by sampling bias. However, some of the correlations are not what would be expected from the sampling bias. For example, there appear to be more young deaths in later years, which is hard to explain by sampling bias.

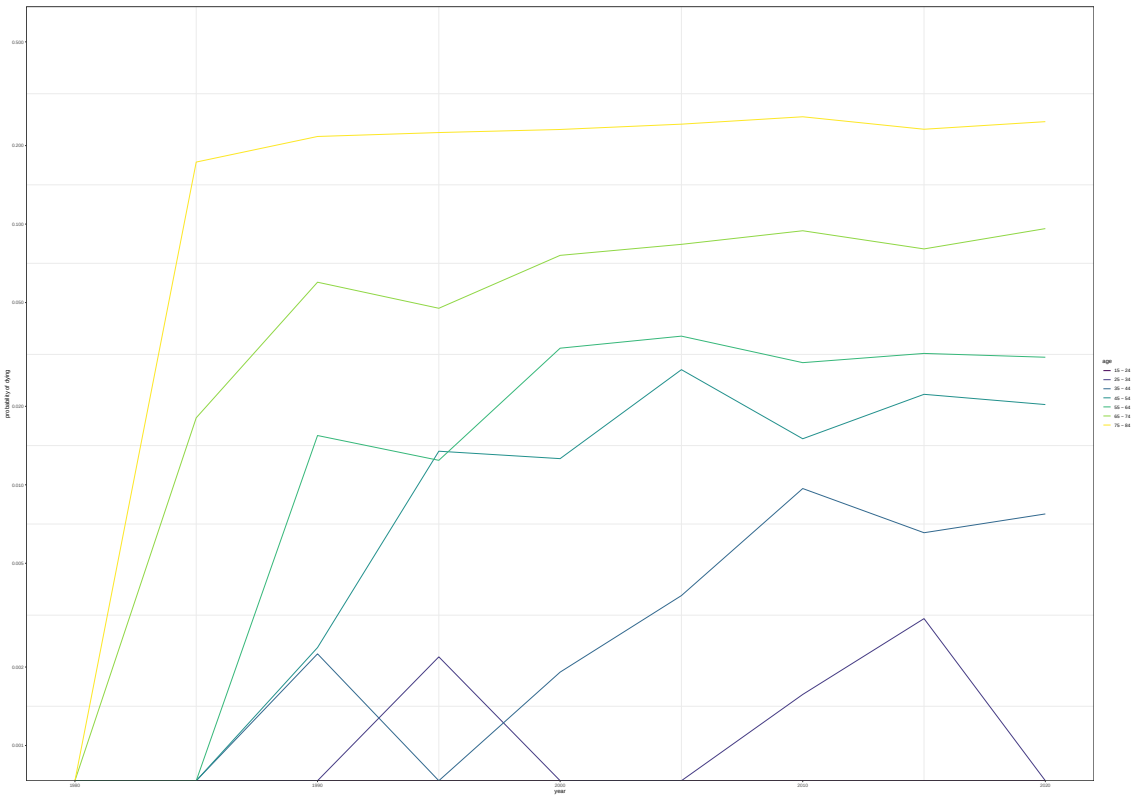
To better explore the data, it would be useful to know the ages of pension beneficiaries at each year, and which of them survive or die.



As in Question 2, the colour indicates the proportion of the individuals at that age who die, while the transparency indicates the sample size. The diagonal patterns in alpha that can be seen in the plot are cohort effects — if a particular year has a large number of plan members of a particular age, those members will be aged 1 year more the following year. The patterns on this plot are not clear. It seems that the probability of dying at each age may be increasing over time. To make it easier to generalise, we create larger blocks, by aggregating 5-year intervals, and 10-year age ranges.



From this plot, the overall probability of dying seems stable over time for each age. We can also plot this as a multiple line plots



This plot suggests that the overall probability of dying in each age range is increasing slightly over time. However, assembling the data in this way loses the additional details of each plan member.

Conclusions

- a
- a
- a
- a
- a
- a

5. *A scientist is studying the effect of social habits on microbial communities in the guts of animals. He collects the following data*

Variable name	*Meaning*
species	*The species of animal*
social.type	*The type of social behaviour of the animal*
diet	*Carnivore, herbivore, omnivore*
age	*The age of the animal at sample collection*
wild	*Whether the animal is wild or captive*
Bacteroides	*The percentage of the gut community consisting of the phylum Bacteroides*
Firmicutes	*The percentage of the gut community consisting of the phylum Firmicutes*

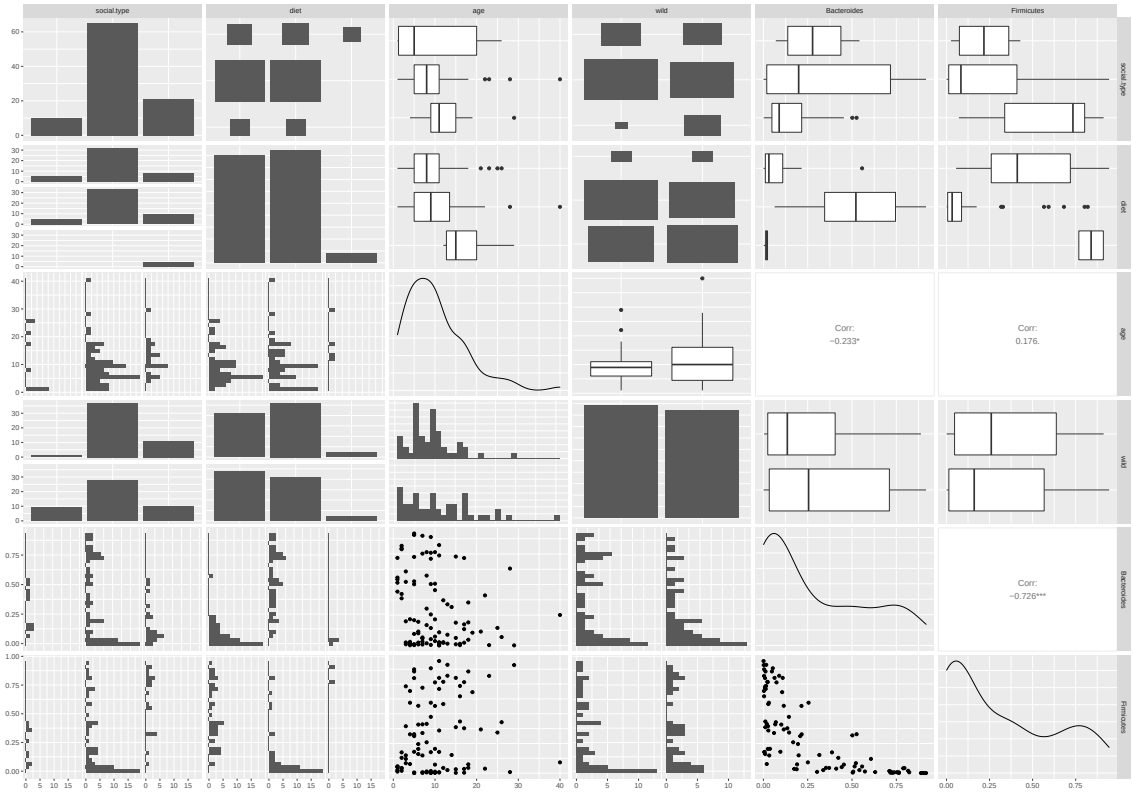
The data are in the file `HW2Q5.txt`*. For captive animals, the species, age and social behaviour are identified by careful examination. For wild animals, they are determined by video surveillance of the habitat, with animals observed fewer than 3 times removed from the sample. The percentages of Bacteroides and Firmicutes are determined by sequencing the bacterial community in a faecal sample. For captive animals, this sample is collected within two hours, and sequenced the following day. For wild animals, the sample is collected at the following site visit, which can be up to a week after the sample is produced. The sample is then sequenced upon return to the laboratory, which may be another week.*

Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.

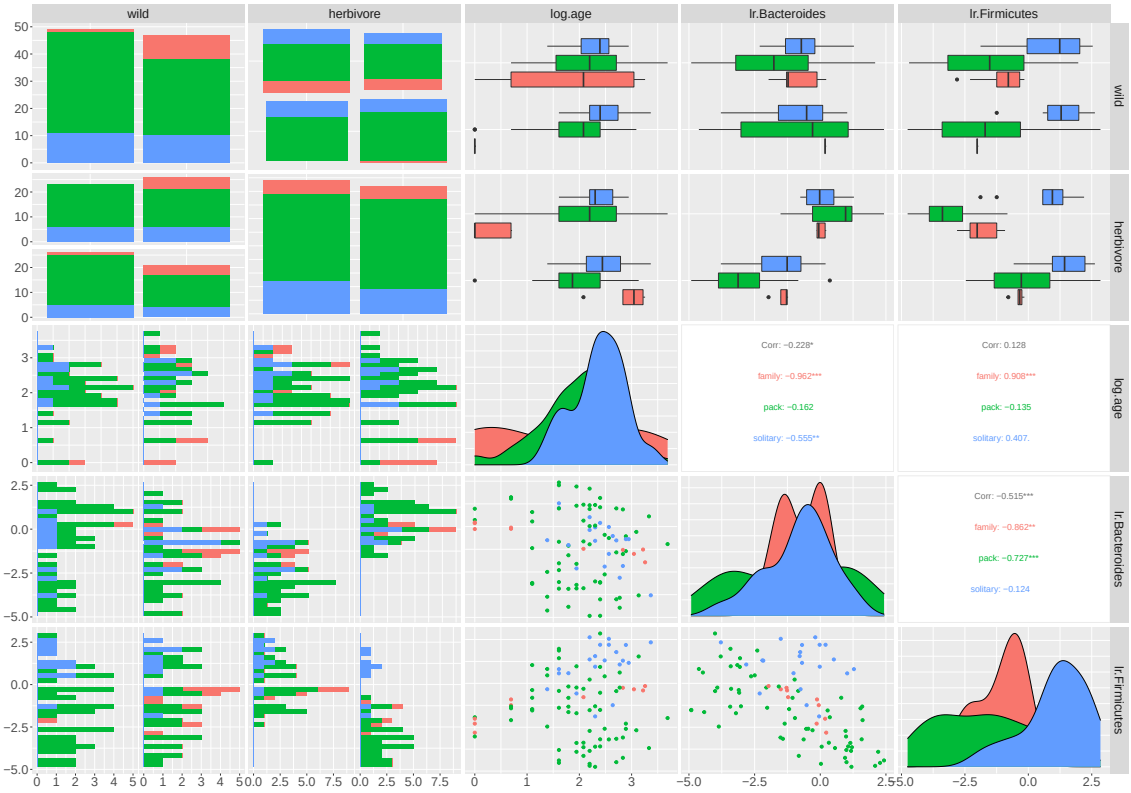
There are obvious data issues with the differences in data collection for wild and captive animals. These could lead to bias in the results. I do not have the expertise to assess how reliable the examination is for determining species and age. There could be some bias here, with age being determined more accurately for some species than others.

We begin with a summary of the data and pairwise scatterplots.

```
  species     social.type       diet            age            wild         Bacteroides       Firmicutes      
lion    :10   family  :10   carnivore:45   Min.   : 1.00   Mode :logical   Min.   :0.0010   Min.   :0.00100  
seal    : 9   pack    :65   herbivore:47   1st Qu.: 5.00   FALSE:     49   1st Qu.:0.0275   1st Qu.:0.03525  
deer    : 7   solitary:21   omnivore : 4   Median : 9.00   TRUE :     47   Median :0.1810   Median :0.19200  
horse   : 7                                Mean   :10.09                   Mean   :0.2947   Mean   :0.31619  
pig     : 7                                3rd Qu.:13.25                   3rd Qu.:0.5230   3rd Qu.:0.58475  
dog     : 6                                Max.   :40.00                   Max.   :0.9180   Max.   :0.94400  
(Other):50                                                                                                   
```

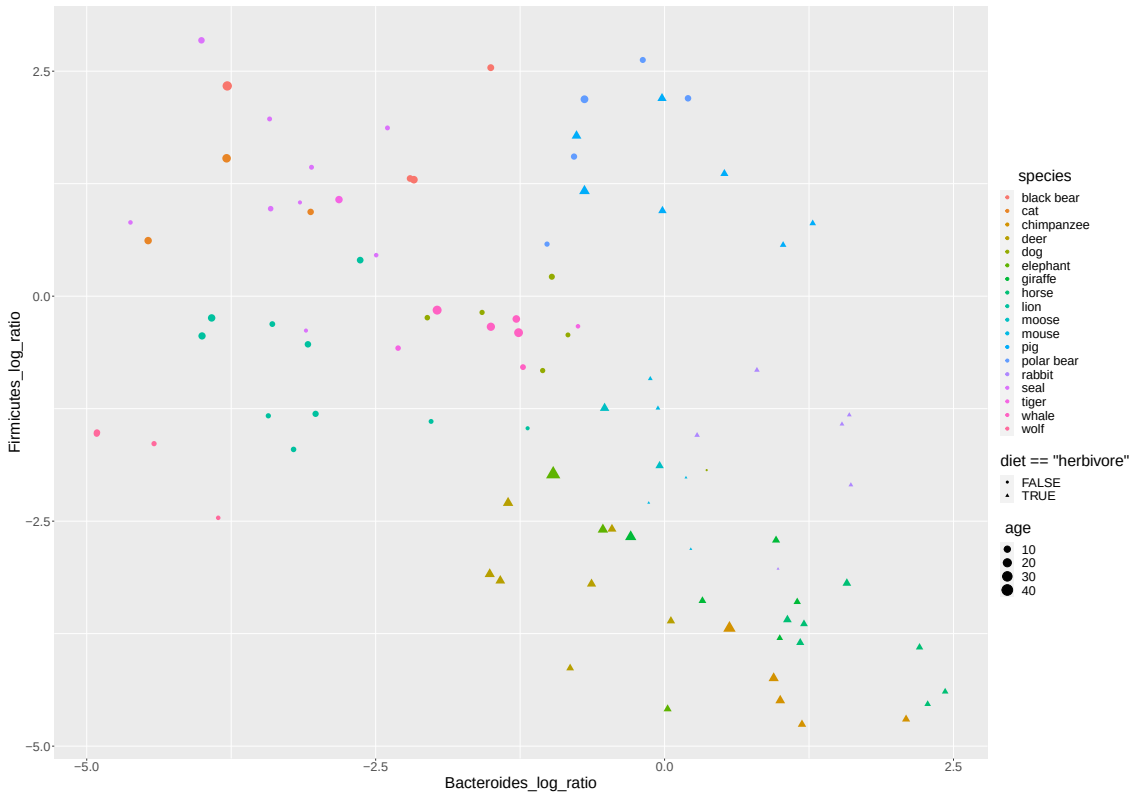


There are too many species, and not enough observations of each species to include in the scatterplots. Otherwise, the data seems fairly clean, with no obvious outliers. The three continuous variables are all positively skewed. There are a few possible outliers in age, but this could simply be the skewed distribution. The compositionality between Bacteroides and Firmicutes (the fact that they are mutually exclusive proportions, so must add up to at most 1) clearly induces a negative correlation between them. A number of factor variables are unbalanced. For example, there are only 4 omnivores in the data set. A brief glance at the pairwise scatterplots suggests that the omnivores are more similar to carnivores than herbivores. Since the data set is small, we will combine the omnivores and carnivores, instead of removing the omnivores. It is also appropriate to log-transform age. For the Bacteroides and Firmicutes variables, it is also possible to log-transform them, but less appropriate because of the constraints. An alternative transformation is to take logarithms of the ratios $\frac{\texttt{Bacteroides}}{1-\texttt{Bacteroides}-\texttt{Firmicutes}}$ and $\frac{\texttt{Firmicutes}}{1-\texttt{Bacteroides}-\texttt{Firmicutes}}$, where the remaining bacteria are used as a reference class. We perform these transformations and redraw the scatterplots, using `social.type` for the colour.

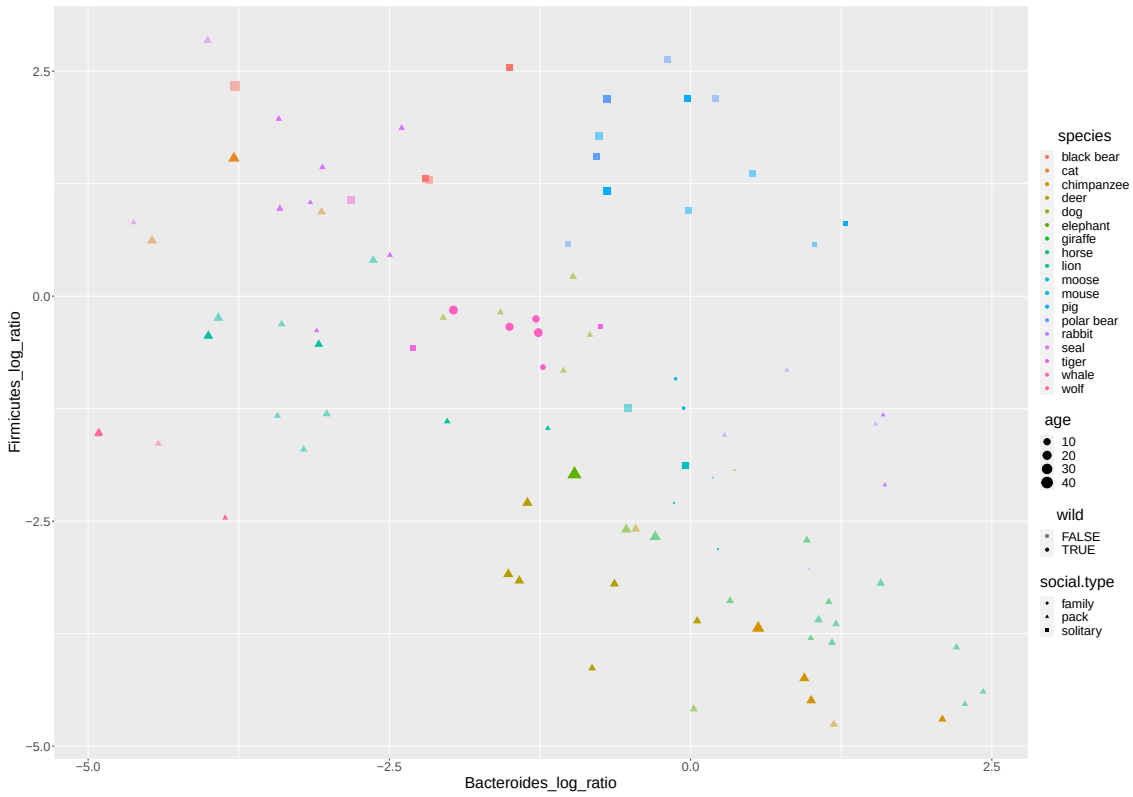


The transformed variables mostly have more normal distributions. There is still some correlation between the transformed Bacteroides and Firmicutes, but this may be because of the compositionality.

As the bacterial composition is the main variable of interest, we will plot the transformed proportions on the x and y axis. We can then use colour, size and shape to indicate the other variables.



We see that animals from the same species are clustered together, and that herbivores have a higher relative abundance of Bacteroides, while carnivores have a lower relative abundance of Bacteroides.



We see that social.type, age and wild do not have a large effect on the microbial community.

Conclusions

- The sampling could produce bias in the data, with the most challenging being the difference in data collection between wild and captive animals.
- The data are fairly clean, without any obvious ouliers or incorrect values.
- The numerical values are positively skewed, and it may be better to transform them.
- There is a large negative correlation between Bacteroides and Firmicutes because of the compositionality.
- Animals from the same species have similar gut communities.
- Herbivores have more Bacteroides than carnivores and omnivores.
- Captivity, age and social behaviour do not appear to have clear effects on the microbial community.