ACSC/STAT 3740, Predictive Analytics

WINTER 2025

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.]

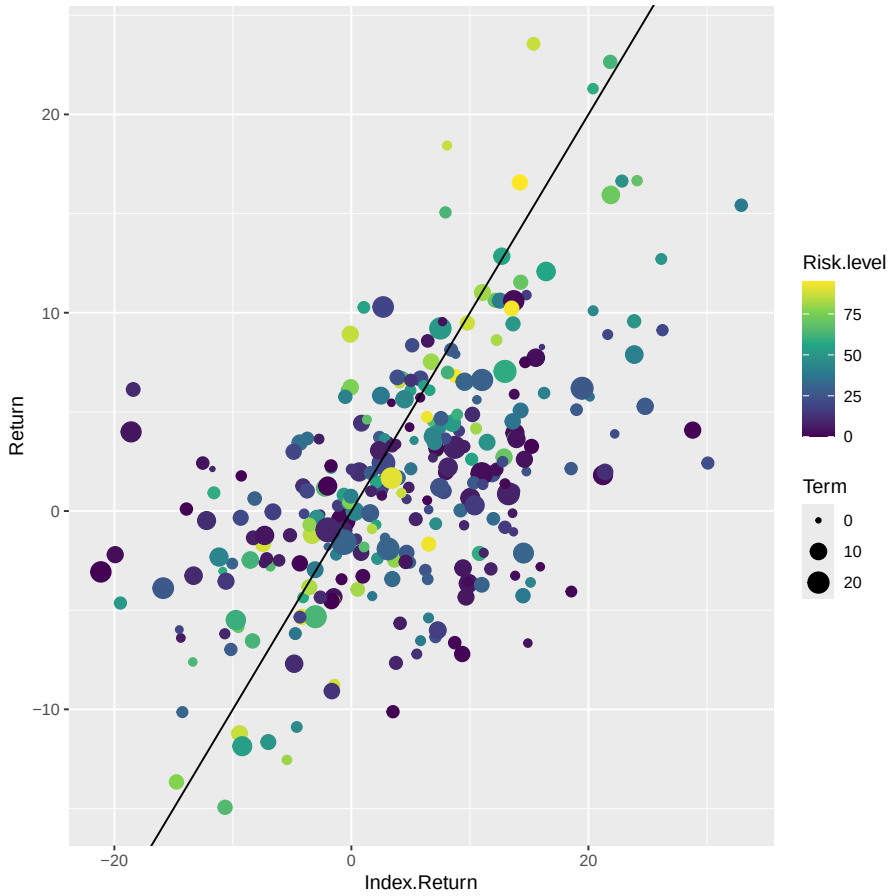
Standard Questions

1. *The file* `HW2Q1.txt` *contains the following data from an insurance company's records on investment returns.*

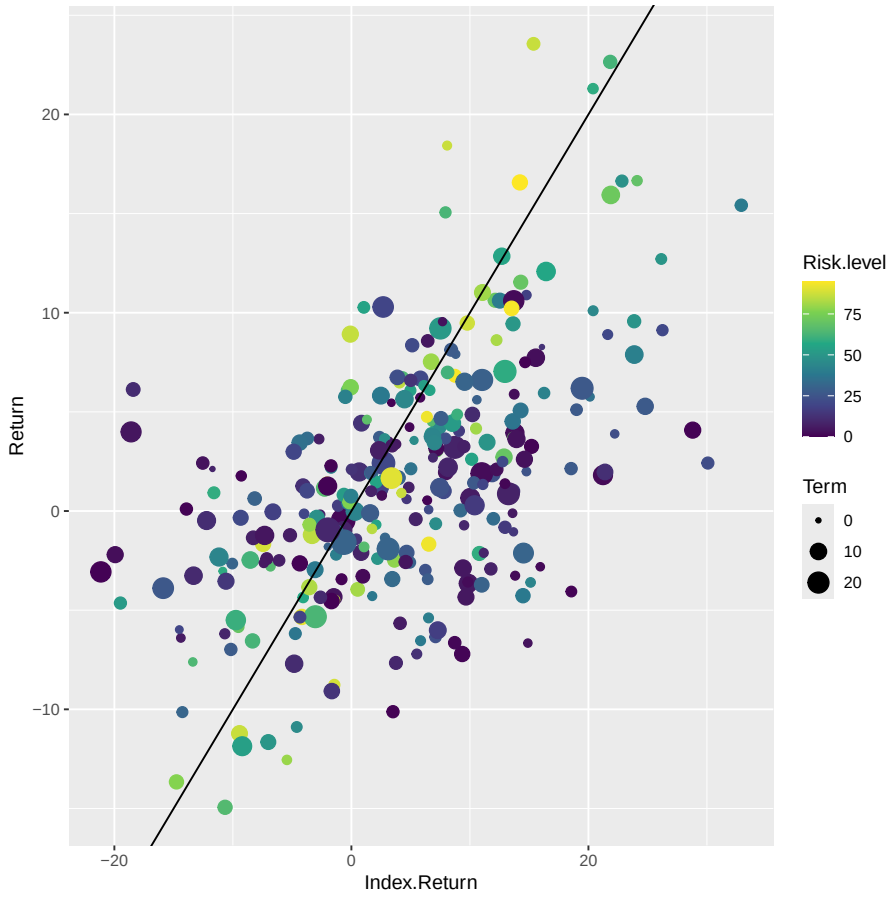
Variable	*Meaning*
Term	*The length of time the investment was to be held*
Liquidity	*A measure of the liquidity of the investment*
Risk.level	*A measure of the relative risk of the investment*
Index.Return	*The return on a comparable market index.*
Return	*The percentage return on the investment.*

Construct a plot or plots to show this data for the purpose of data exploration.

The following simple plot shows most of the information.



An alternative approach is to plot the ratio $\frac{\texttt{Return}}{\texttt{Index.Return}}$. However, this has some large outliers when the index return is close to zero. Restricting to points where the absolute value of the index return is more than 5 gives the following plot:



This plot allows us to show more information, and gives some interesting patterns, but does omit some of the data.

The plots were created using the following code.

```
HW2Q1<-read.table("HW2Q1.txt")

### Plot a - straightforward scatterplot

ggplot(HW2Q1,mapping=aes(x=Index.Return,
                         y=Return,
                         size=Term,
                         colour=Risk.level))+
    geom_point()+
    geom_abline()+
    scale_colour_viridis_c()



### Plot b - plot Return/Index.Return

ggplot(HW2Q1[HW2Q1$Index.Return^2>25,],mapping=aes(x=Risk.level,
                                                   y=Return/Index.Return,
                                                   size=Term,
                                                   colour=Liquidity))+
    geom_point()+
    scale_colour_viridis_c()+
    scale_y_continuous(limits=c(-1.2,2.5))
```

2. *The file* **`HW2Q2.txt`** *contains the following data from an experiment on the effect of climate on fertility of wolves.*

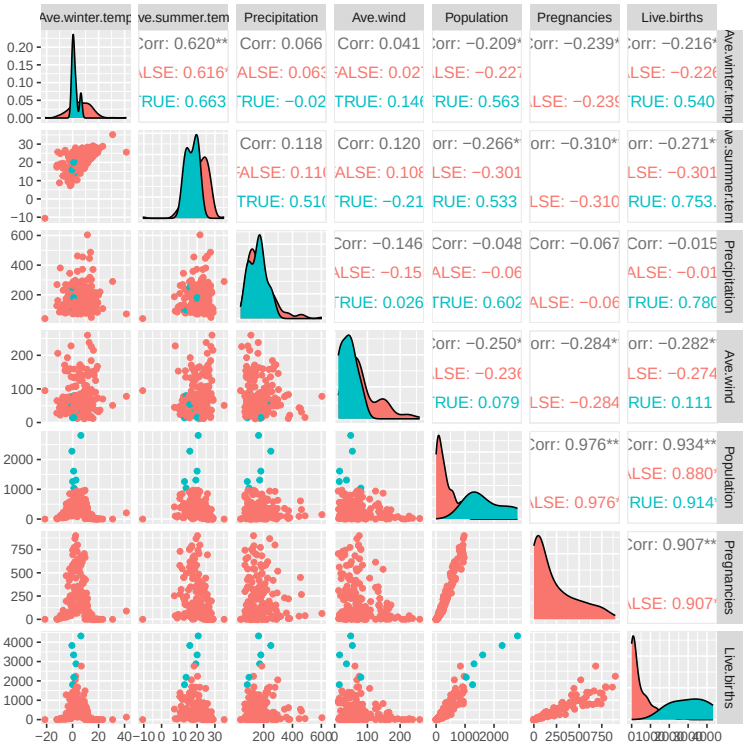
Variable	*Meaning*
Ave.winter.temp	*The average daily maximum temperature in the period Dec–Mar*
Ave.summer.temp	*The average daily maximum temperature in the period Jun–Aug*
Precipitation	*The total annual precipitation*
Ave.wind	*The average wind speed during the year.*
Population	*The total adult population of the pack.*
Pregnancies	*The number of pregnancies.*
Live.births	*The number of live births in the pack.*

The climate data are average readings from a nearby weather station over the previous 10-year period. The population, pregnancies an live births data are estimated using a capture-recapture experiment, where wolves are marked, and then set loose, and the proportion of observed individuals marked is used to estimate the total population.

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 consider the data collection process. Readings from the weather station should be fairly accurate, but could have some small bias. The capture recapture experiment could be a bigger source of bias. The estimated population is based on the assumption that all wolves have the same chance of being captured. If this assumption is not correct, then the populations will be systematically underestimated, which could cause bias in the analysis.

We start by summarising the data and plotting pairwise scatterplots. The summary immediately shows that there are 6 NA values for Pregnancies. We highlight these points on our pairwise scatterplots. We see that they corresond the to the largest 6 populations.



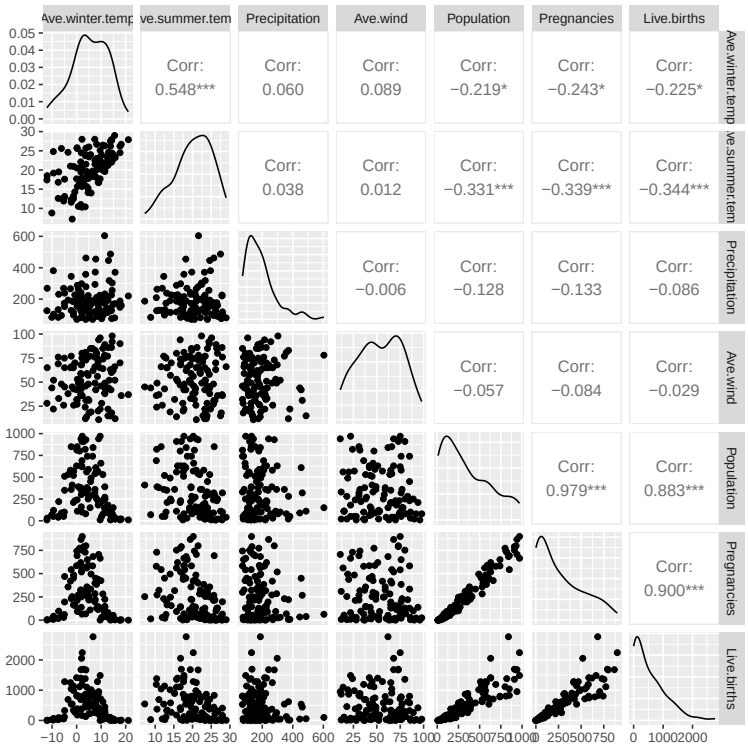
These missing values are therefore, not completely at random, and removing them may cause bias for estimating the relationships between some variables. However, as these values are outside the range of the complete data, imputing the missing values could involve a dangerous amount of extrapolation. Thus removing the incomplete observations is probably the best approach.

These scatterplots highlight a number of outliers, some of which are implausible. For example, there are some sites where average winter temperature is higher than average summer temperature. There is one site where the average summer temperature is around $-10°C$. This is possible, but might be removed. There is an outlier with average summer temperature about $35°C$, which is plausible, but is far enough from other values that it might be too influential or not follow the pattern of other sites. There are also some very large values for precipitation, which are possible, but might be too influential. Many values for average wind are implausible, including many over 100 Kph.

We also check for duplicated values. We find that there are three du-

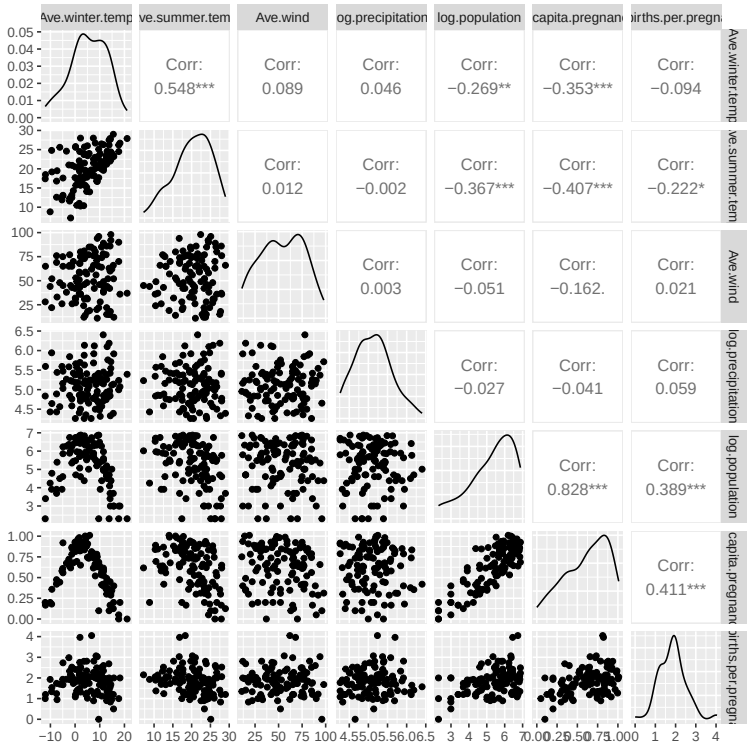
plicated values. Given the large number of numerical variables, it seems implausible that these duplicates could be entirely by chance, so we remove them.

After removing the implausible values and outliers, we replot the scatterplots.



We see that there is a very strong relation between population and pregnancies, and between pregnancies and live births. This suggests creating the features per-capita pregnancies and live births per pregnancy, which may be more informative. An alternative modelling approach would be a Poisson regression with population as an offset.

We also note that population and precipitation are heavy-tailed, so log-transformation might be appropriate. After these transformations, we replot the scatterplots.



We see that there is a strong non-linear relation between average winter temperature and both log population and per-capita pregnancies. There appears to also be a weaker linear relation between log population and per-capita pregnancies.

In summary, we have the following conclusions from the data exploration:

- The capture-recapture experiment cause bias, potentially leading to the populations being underestimated.
- There are 6 missing values for pregnancies. These correspond to the largest values for population. I have removed these observations, but this could lead to bias in the results.
- Entries 35, 100, and 102 are duplicates of entries 34, 99, and 101 respectively. These entries have many numerical values, so it is unlikely that the duplicates could have occured by chance. I have therefore removed the duplicated entries.
- One site has higher average winter temperature than summer temperature. This seems wrong, so we remove this site.
- One site has an outlier in summer temperature, which is not implausible, but might be removed anyway, in case it is too influential.
- Precipitation also has a very heavy-tailed distribution, so should probably be log-transformed.

- There are a number of sites with implausible average wind speeds. I have removed speeds over 100 kph, but other choices of cut-off are reasonable, or we could exclude the whole wind speed variable.
- There are some sites with population=0. While these are possible, they are not relevant to the analysis, so I have removed them.
- There is a clear relation between population, pregnancies and live births, so I have created the variables per-capita pregnancies, and live births per pregnancy.
- Population is heavy-tailed, so I have log-transformed it.
- After the transformations, there are strong relationships between average winter temperature, population, per-capita pregnancies and live births per pregnancy. The relationships between log population, per-capita pregnancies, and live births per pregnancy are approximately linear. The relationships with average winter temperature is nonlinear.
- Average summer temperature, wind speed and precipitation do not show strong associations with population, pregnancies or live births.
- As the response is count data, Poisson regression, or overdispersed Poisson regression may be a good approach. A natural choice would be to use population as an offset for pregnancies, or pregnancies as an offset for live births.

The following code was used for this exploration.

```
HW2Q2<-read.table("../HW2Q2.txt")
library(GGally)


summary(HW2Q2)
ggpairs(HW2Q2,mapping=aes(colour=is.na(HW2Q2$Pregnancies)))

which(duplicated(HW2Q2))
HW2Q2[duplicated(HW2Q2),]

HW2Q2_good<-HW2Q2[!duplicated(HW2Q2),]%>%
    filter(
        Population>0,
        Ave.winter.temp<Ave.summer.temp,
        Ave.summer.temp<34,
        Ave.wind<100,
        !is.na(Pregnancies))

ggpairs(HW2Q2_good)

ggpairs(HW2Q2_good%>%mutate(
                        "log.precipitation"=log(Precipitation),
                        "log.population"=log(Population),
                        "per.capita.pregnancies"=Pregnancies/Population,
                        "live.births.per.pregnancy"=Live.births/Pregnancies
                    )%>%select(
                            -c("Live.births","Pregnancies","Population",
                               "Precipitation")))
```

3. *A government has collected the following data about the effect of educational grants on social mobility in the file* `HW2Q3.txt`.

Variable	*Meaning*
GDP.growth	*The annual growth in GDP.*
Gini.coefficient	*A measure of income inequality in the country.*
Political.system	*The system of government in the country.*
Percent.Tech	*The percentage of GDP attributable to the technology industry.*
Percent.Service	*The percentage of GDP attributable to the service industry.*
Percent.Manufacture	*The percentage of GDP attributable to the manufacture industry.*
Percent.Agriculture	*The percentage of GDP attributable to the agriculture industry.*
Percent.Resources	*The percentage of GDP attributable to the resources (e.g. mining) industry.*
Unemployment	*The percentage of individuals seeking employment who are unable to find it.*
Education.years	*The average number of years spent in full-time education.*
Percent.University	*The percent of individuals who attend university.*
Education.Grants	*The per-capita amount spent on educational grants.*
Social.Mobility	*An index measuring social mobility.*

The economic data are from the government websites for each country. Political data are from the classification of government systems in an academic paper. Data on education systems, university attendance are from a website giving international survey results on education. Education grant data are obtained from government websites. Social mobility data are from an international website that provides survey results about social mobility and various other lifestyle factors.

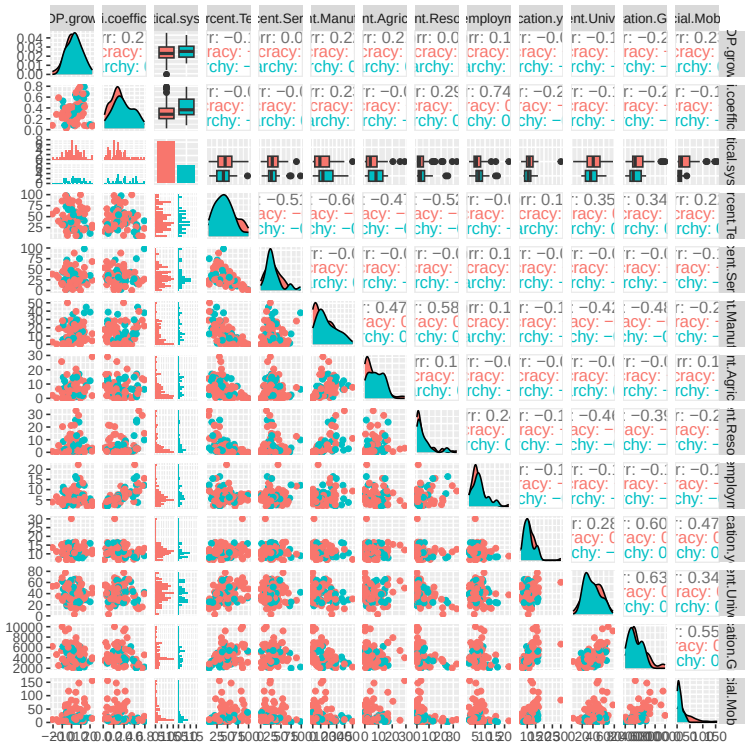
Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

Considering the data collection, there is certainly some possibility of bias here. Government websites may not be reliable, and the bias may be different for each country. We would need more information on the classification methodology from the academic paper, but it is probably fairly well correlated with the desired classification. Survey results can be influenced by the public information available, which may vary between countries.

Looking at the data, we first note that there are 34 NA values for education grants. Checking these values, we see that they are exactly the autocracies. This means that removing them will limit the analysis to other political systems, but may provide a good analysis for those systems. I have therefore taken that approach. Another possibility might be to remove the variable `Education.Grants`.

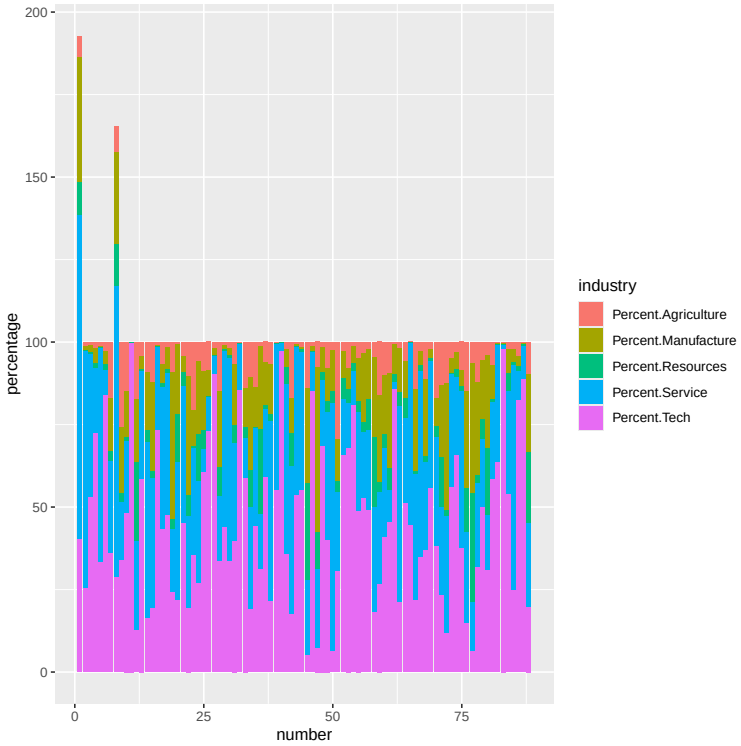
Checking for duplicates, we see that records 22 and 23 and records 118 and 119 are duplicates. These are unlikely to be genuine duplicates, given the number of numerical variables, particularly given that they are consecutive. We therefore remove the duplicates.

We next plot pairwise scatterplots of the complete cases. We colour by political system:

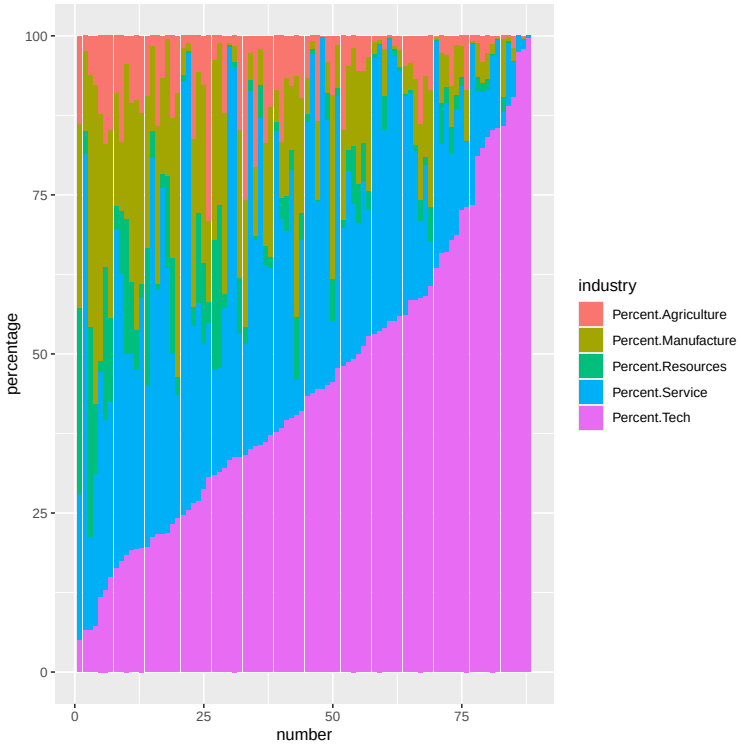


We see that GDP growth and Gini coefficient are fairly normally distributed and quite correlated. There are three outliers with more negative GDP growth than other countries and there are five outliers with high GDP growth and very low Gini coefficient. On their own, these would not be outliers, but given the relation between GDP growth and Gini coefficient for other countries, these are outliers. We examine these in more detail: for both groups of outliers, there is nothing very obvious about the data to suggest anything incorrect. I would suggest checking the values and analysing both with and without these points to determine how influential they are.

The percent of the economy attributed to different industries has a fairly normal distribution for tech and services, and a slightly skewed distribution for manufacture, agriculture and resources. This might be better viewed as as stacked bar-chart. We could use the x-axis to display some relevant information, but in the first plot, we use it for row number.



This plot immediately highlights a problem. The percentages do not always add up to 100. We see that there are 2 observations where the percentages do not add up to 100 ignoring rounding errors. In both cases, the discrepancy could be explained by a single error in `Percent.Service`, which is extremely high for both countries. We should check this value if possible. For the initial exploration and modelling, we will correct the values of `Percent.Service`. After this correction, we get the following stacked bar-chart, which we have ordered by `Percent.Tech` in order to improve readability.



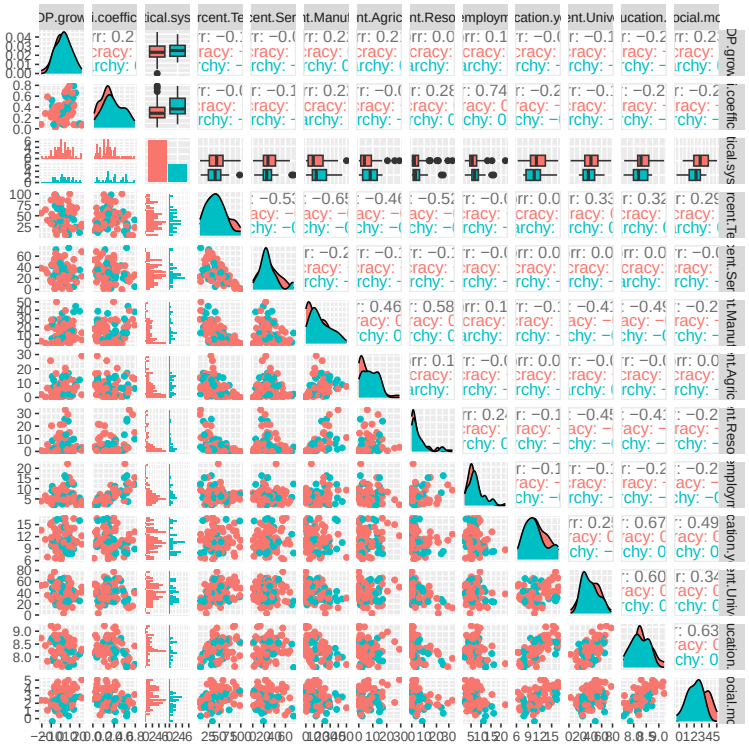
It seems there is quite a lot of variance in the composition of each countries economy, but no really clear outliers. There are some transformations that we might consider for the percentages of each industry, such as a log-transformation. For the initial analysis, it does not seem necessary.

We see that there are several outliers in `Unemployment` and one outlier in `Education.years`. The outlier in `Education.years` is extremely implausible, so we remove it. The outliers in unemployment are plausible. I would suggest comparing the analysis results with and without them.

`Education.Grants` seems slightly skewed, which might either be fixed by removing a few outliers or with a log-transformation.

We see that social mobility is low in monarchies and has a heavy-tailed distribution for democracies. There also appear to be several outliers. We therefore perform a log-transformation, and see whether these remain outliers after the transformation.

After removing the outlier in `Education.years` and log-transforming `Education.Grants` and `Social.Mobility`, we replot the pairwise scatter-plots.



We see that political system is strongly associated with social mobility. Of the different industries, percent tech shows a significant positive correlation with social mobility, while percent manufacture and percent resources are negatively correlated with social mobility. The percent of individuals who attend university, years of education and education grants all show strong positive correlation with log-transformed social mobility. GDP growth shows weaker positive correlation with social mobility. Gini coefficient and unemployment are negatively correlated with log-transformed social mobility, particularly in monarchies.

Among the predictors, there is a strong positive linear relation between education years, education grants and percentage of individuals who attend university.

We conclude the following from our data analysis:

- The data are from a variety of different sources, many of which may have different biases.
- The amount spent on educational grants is missing for autocracies. I have removed these observations.
- There are two duplicated records, which are clearly not correct, so have been removed.

- Two records have percentages exceeding 100. These have large values for `Percent.Service`, so I have assumed this is the incorrect value, and corrected it, assuming percentages sum to 100.
- Education grants and Social mobility have skewed distributions. I have log-transformed them, after which the distributions are approximately normal.
- There are several outliers in GDP growth and Gini coefficient. I left these outliers in the data, but should compare results with and without the outliers.
- There is one implausible outlier in `Education.years` which I have removed.
- Most relations seem linear.
- Democracy, Percent tech, Education years, Percent University and Education grants are significantly positively correlated with social mobility.
- Gini coefficient, Percent Manufacture, Percent Resources and Unemployment are significantly negatively correlated with social mobility.
- There is a strong positive linear relation between `Education.years`, `Percent.University` and `Education.grants`.

The `R` code used for this data exploration is the following

```
HW2Q3<-read.table("HW2Q3.txt",stringsAsFactors=TRUE)
summary(HW2Q3)
summary(HW2Q3[is.na(HW2Q3$Education.Grants),])
which(duplicated(HW2Q3))
HW2Q3[c(23,119),]
HW2Q3[HW2Q3$Gini.coefficient==0.346,]
HW2Q3[HW2Q3$Gini.coefficient==0.292,]

### Remove duplicates and NA values
library(dplyr)
HW2Q3_good<-HW2Q3[-c(23,119),]%>%filter(!is.na(Education.Grants))

### Make pairwise scatterplots
library(GGally)
ggpairs(HW2Q3_good,mapping=aes(colour=HW2Q3_good$Political.system))

### Examine outliers
HW2Q3_good%>%filter(GDP.growth< -12)
HW2Q3_good%>%filter(GDP.growth>10,Gini.coefficient< 0.2)


### Make stacked barplot of percentages in each industry
library(tidyr)
ggplot(HW2Q3_good%>%
       mutate(number=seq_along(HW2Q3_good$GDP.growth))%>% # create row numbers
       pivot_longer(cols=c(Percent.Tech,  ### Change to long format
                           Percent.Service,
                           Percent.Manufacture,
                           Percent.Agriculture,
                           Percent.Resources),
                    names_to="industry",values_to="percentage"),
       mapping=aes(y=percentage,fill=industry,x=number))+
    geom_col(position="stack")


### Examine outliers
HW2Q3_good[c(1,8),]

### Adjust Percent Service so that totals are 100%
HW2Q3_fixed<-HW2Q3_good
HW2Q3_fixed$Percent.Service<-100-HW2Q3_fixed$Percent.Tech-
    HW2Q3_fixed$Percent.Agriculture-
    HW2Q3_fixed$Percent.Resources-
    HW2Q3_fixed$Percent.Manufacture

### Make new stacked bar-plot ordered by Percent.Tech

ggplot(HW2Q3_fixed%>%
       mutate(number=rank(HW2Q3_fixed$Percent.Tech+runif(dim(HW2Q3_fixed)[1])*0.1))%>% # Add random noise to break ties.
       pivot_longer(cols=c(Percent.Tech,
                           Percent.Service,
                           Percent.Manufacture,
                           Percent.Agriculture,
                           Percent.Resources),
                    names_to="industry",values_to="percentage"),
       mapping=aes(y=percentage,fill=industry,x=number))+
    geom_col(position="stack")

### Make pairwise scatterplots with transformed variables.
ggpairs(HW2Q3_fixed%>%filter(Education.years<20)%>%
        mutate(log.education.grants=log(Education.Grants),
               log.social.mobility=log(Social.Mobility))%>%
        select(-c("Education.Grants","Social.Mobility")),
        mapping=aes(colour=Political.system))
```

4. *The file* `HW2Q4.txt` *contains the following data from a company's human resources department.*

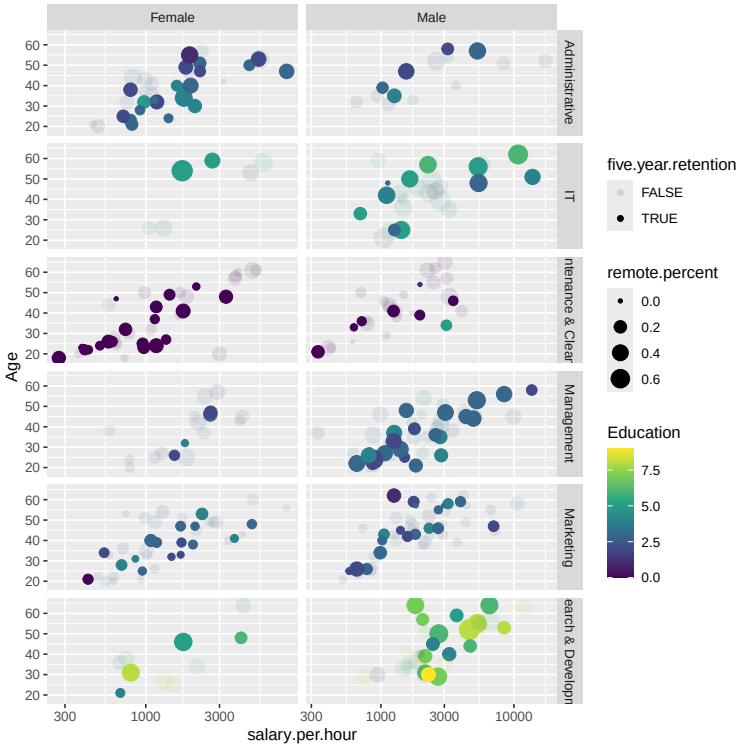
Variable	*Meaning*
Age	*The employee's age.*
Gender	*The employee's gender.*
Education	*The number of years of post-secondary education that the employee has.*
Job.type	*The type of job the employee has.*
Salary	*The employee's annual salary.*
Total.hours	*The employee's average number of weekly hours.*
Remote.hours	*The employee's average number of hours working remotely.*
5-year retention	*Whether the employee remains at the company for 5 years.*

The data coming from the companies own records should be fairly reliable. Any missing data may not be missing at random. The education data may be reported by the employee, and might not be independently verified, so there could be inaccuracies.

A quick summary of the data reveals that there are 26 missing values for `Remote.hours`. These correspond to the individuals aged over 65. We also see that some of the ages are completely implausible. It therefore makes sense to remove these missing values, and restrict to ages below 65. We make pairwise scatterplots coloured by gender.

We see that there are two clear outliers in `Education`. These values are not impossible, but should probably be removed. We see `Salary` has a skewed and heavy-tailed distribution, which suggests applying a log-transformation. `Remote.hours` also has a skewed distribution, and might benefit from log-transformation. Two obvious features that could be added to the dataset are salary per hour and percentage of remote hours. [Technically salary per hour is not correct as hours are per-week and salary is per year. For data exploration, this is probably not a serious problem.] After removing the outliers in education, creating the new features and log-transforming the skewed features, we get the following pairwise scatterplots:

From this plot, we see that `Gender` is related to a lot of other variables. We also see that we have removed most of the outliers. We can actually fit a lot of the information onto a single plot. For example:



This figure shows several patterns. There is a clear linear relation between age and log-salary per hour for each gender and job type. Education level differs a lot between job types. There is not an obvious relation between the predictors and retention. For the categorical predictors, we can make a table

Job.type	Female		Male	
	retained	left	retained	left
Administrative	21	14	5	9
IT	2	4	11	15
Maintenance & Cleaning	19	18	8	22
Management	4	13	22	24
Marketing	16	23	19	30
Research & Development	4	6	15	16

This shows that retention rate varies by gender and job type, and possibly with the interaction of the two variables. There is not a clear pattern for which employees are retained for each gender and job type.

We have reached the following conclusions from our data exploration:

- There are 26 missing values for `Remote.hours`, corresponding to the individuals aged over 65. There are also a number of implausible ages, so I have removed these data points.

- There are no duplicated records.
- There are two outliers in `Education`, which I have removed.
- I have added two features: Salary per hour and percentage of remote hours.
- Salary per hour has a skewed heavy-tailed distribution, so I have log-transformed it.
- The interaction between Gender and Job type is an important predictor of 5-year retention
- Within each gender and job title, there is a strong linear relation between age and salary per hour.
- There is not an obvious relation between the other predictors and 5-year retention.

The `R` code used for this data exploration is the following

```
HW2Q4<-read.table("HW2Q4.txt",stringsAsFactors=TRUE)
summary(HW2Q4)
library(dplyr)
### Summarise the missing values
summary(HW2Q4%>%filter(is.na(Remote.hours)))
summary(HW2Q4%>%filter(!is.na(Remote.hours)))

which(duplicated(HW2Q4)) # Check for duplicates.

HW2Q4_clean<-HW2Q4%>%filter(!is.na(Remote.hours))

library(GGally)
### Make pairwise scatterplots coloured by gender.
ggpairs(HW2Q4_clean%>%select(-c("Gender")),
        mapping=aes(colour=HW2Q4_clean$Gender))
### Remove outliers
HW2Q4_main<-HW2Q4_clean%>%filter(Education<10)
### Create new features and log transform skewed heavy-tailed features
ggpairs(HW2Q4_main%>%mutate(log.salary.per.hour=log(Salary/Total.hours),
                            log.remote.percent=log(Remote.hours/Total.hours))%>%
        select(-c("Salary","Remote.hours","Gender")),
        mapping=aes(colour=HW2Q4_main$Gender))

### Make a plot to show the interactions.
ggplot(HW2Q4_main%>%mutate(salary.per.hour=Salary/Total.hours,
                           remote.percent=Remote.hours/Total.hours),
       mapping=aes(x=salary.per.hour,y=Age,size=remote.percent,
                   colour=Education,alpha=five.year.retention))+
    geom_point()+
    facet_grid(Job.type~Gender,scale="free_x")+
    scale_x_log10()+
    scale_colour_viridis_c()
```

5. *An advertising company is studying internet search terms. It collects the following data :*

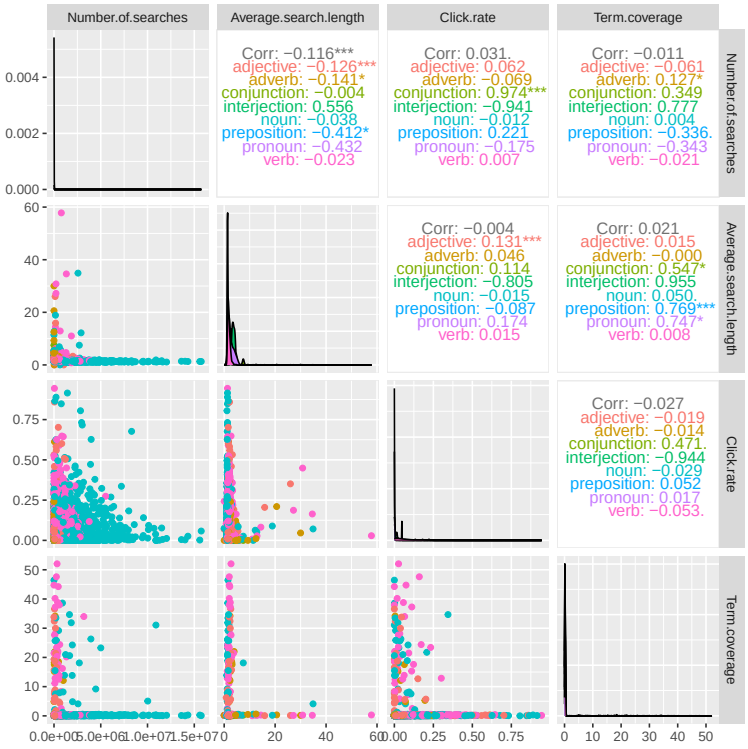
Variable	*meaning*
Part.of.speech	*The grammatical type of word that the search term is.*
Number.of.searches	*The number of searches involving this search term.*
Average.search.length	*The average length in words of a search involving this term.*
Click.rate	*The proportion of searches involving this term that result in a click on an advertisement.*
Term.coverage	*The proportion of searches involving this term that also involve one of the 100 most common search term*

The data are in the file `HW2Q5.txt`.

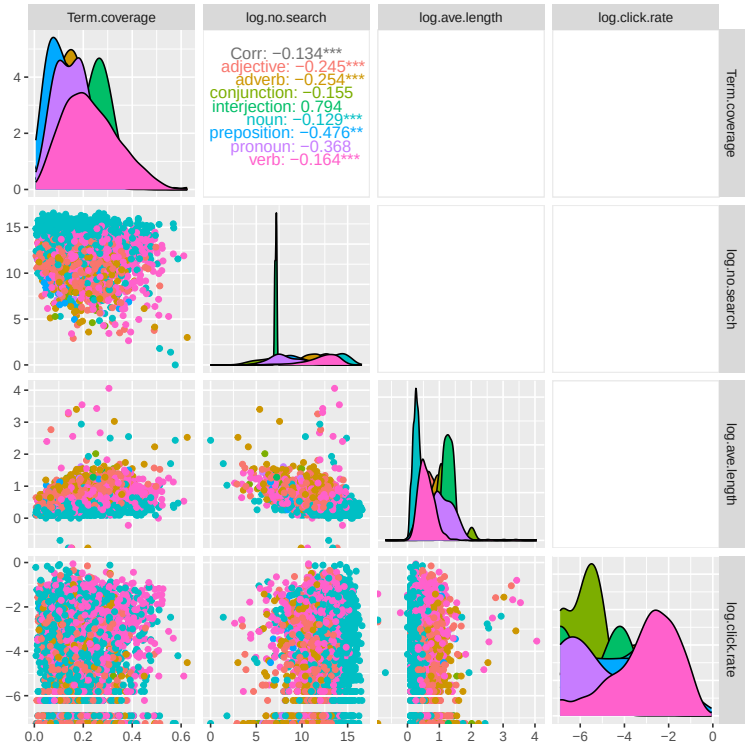
Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models.

We are not given information about how the data were collected, so it is hard to judge the reliability. Given the large nature of the data, it is likely that it comes from some automated source. This should be reliable, but there are questions that might need to be considered. For example, are misspellings included? Also, the sample might be from a limited selection of search engines, which could cause bias in the results if the results are applied to other search engines.

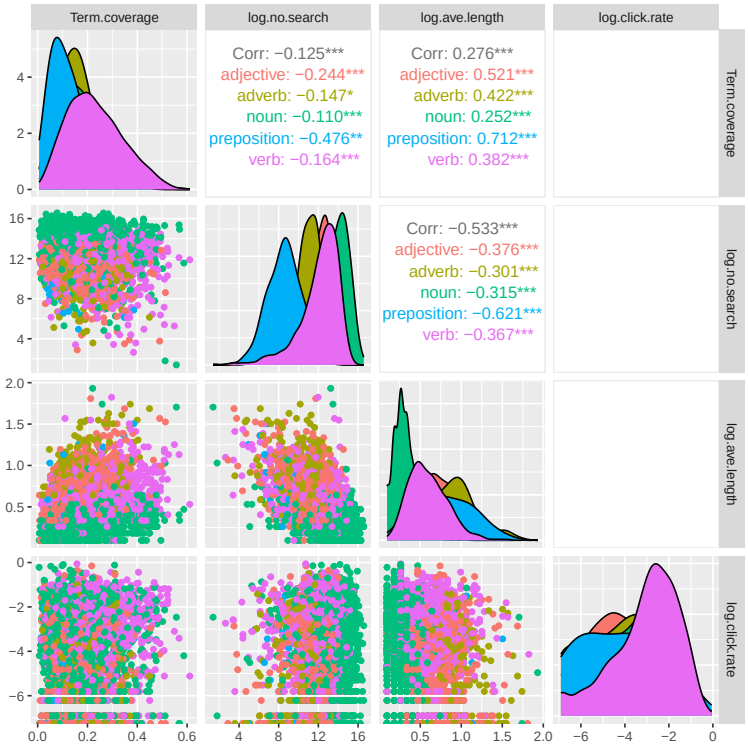
A quick summary of the data reveals that there are 76 missing values, which appear to be missing at random. Given the large size of the data set, it is probably reasonable to remove these observations. We then make pairwise scatterplots coloured by part of speech



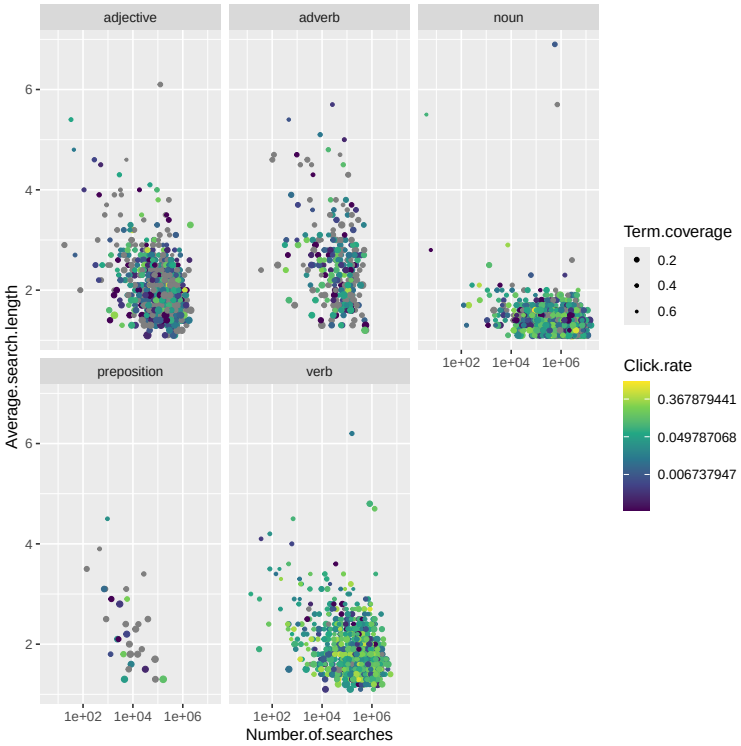
We see that all the predictors are heavy-tailed and skewed. We also note that Term coverage is supposed to be a proportion between 0 and 1, but there are many values greater than 1. It is likely that these values are percentages instead of proportions. Examining these data points, they seem to be random, so the safest approach may be to remove these data points. After removing these points and log-transforming number of searches, average search length and click rate, we make new pairwise scatterplots coloured by part of speech.



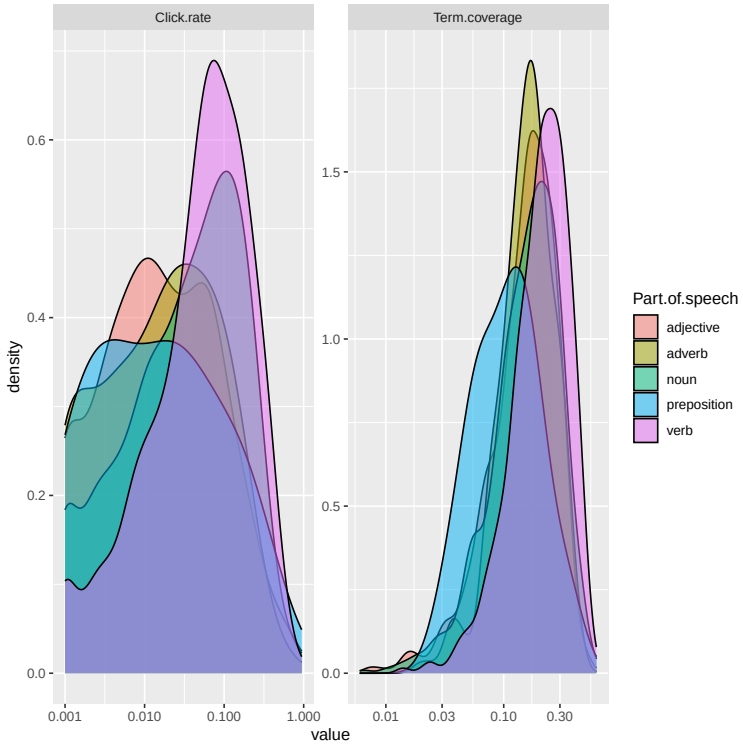
We note that there are a few cases where average search length is less than 1, which seems to be an error. There are also some large outliers for average search length, some of which are implausible. We therefore remove these cases. After the log-transformation, the variables look much more normal for each part of speech, but the distributions are different for each part of speech. The number of terms for each part of speech varies a lot. Given the difference in variance between different parts of speech, it may also make sense to remove the rarer parts of speech. After removing outliers and rarer parts of speech, we replot the pairwise scatterplots.



As there are only a few predictors, it is possible to display them on a single plot.



There is a clear negative association between number of searches and average search length. This association could be linear after log-transforming the number of searches. We do not see a clear pattern for the click-rate or term coverage. Results seem different for different parts of speech. We plot histograms coloured by part of speech.



This shows that nouns and verbs have higher click rates and that prepositions have lower term coverage.

We have identified the following from our data analysis:

- There are 76 missing values for `Click.rate`, which appear to be missing at random. I have removed these values.
- There are no duplicated records.
- There are some values of `Term.coverage` larger than 1, which is impossible. These are probably percentages, but to be safe, I have removed these values.
- Some values of `Average.search.length` are less than 1, which is impossible. Other values are more than 10, which is possible, but implausible. I have removed these values.
- Several parts of speech have very few terms, so it is difficult to reach any firm conclusions. I have removed these cases.
- The numerical predictors have skewed heavy-tailed distributions, so I have log-transformed them.
- There is a clear negative association between `Average.search.length` and `Number.of.searches`. This might be linear on the log-scale.

- Nouns and verbs have higher click rates than other parts of speech
- Prepositions have lower term coverage than other parts of speech.

The data exploration was performed using the following code.

```
HW2Q5<-read.table("../HW2Q5.txt",stringsAsFactors=TRUE)
summary(HW2Q5)
summary(HW2Q5[is.na(HW2Q5$Click.rate),])
summary(HW2Q5[!is.na(HW2Q5$Click.rate),])

### Check for duplicates
sum(duplicated(HW2Q5))

### Remove NAs
HW2Q5_complete<-HW2Q5[!is.na(HW2Q5$Click.rate),]

library(GGally)

ggpairs(HW2Q5_complete[,-1],mapping=aes(colour=HW2Q5_complete$Part.of.speech))

### Check impossible outliers, and remove them
summary(HW2Q5_complete[HW2Q5_complete$Term.coverage>1,])
HW2Q5_good<-HW2Q5_complete[HW2Q5_complete$Term.coverage<=1,]

library(dplyr)
library(GGally)
library(tidyr)

### Log transform long-tailed predictors
ggpairs(HW2Q5_good%>%mutate(log.no.search=log(Number.of.searches),
                            log.ave.length=log(Average.search.length),
                            log.click.rate=log(Click.rate))%>%
        select(-c("Number.of.searches",
                  "Average.search.length",
                  "Click.rate",
                  "Part.of.speech")),
        mapping=aes(colour=HW2Q5_good$Part.of.speech))


### Remove more outliers
HW2Q5_clean<-HW2Q5_good%>%filter(Average.search.length>1&
                                 Average.search.length<7)


summary(HW2Q5_clean)

### Remove rare parts of speech
HW2Q5_cleaned<-HW2Q5_clean%>%filter(Part.of.speech%in%c("noun",
                                                        "verb",
                                                        "adjective",
                                                        "adverb",
                                                        "preposition"))


ggpairs(HW2Q5_cleaned%>%mutate(log.no.search=log(Number.of.searches),
                               log.ave.length=log(Average.search.length),
                               log.click.rate=log(Click.rate))%>%
        select(-c("Number.of.searches",
                  "Average.search.length",
                  "Click.rate",
                  "Part.of.speech")),
        mapping=aes(colour=HW2Q5_cleaned$Part.of.speech))


ggplot(data=HW2Q5_cleaned,
       mapping=aes(x=Number.of.searches,
                   y=Average.search.length,
                   size=Term.coverage,
                   colour=Click.rate))+
    geom_point()+
    facet_wrap(Part.of.speech~.)+
    scale_x_log10()+
    scale_colour_viridis_c(trans="log")+
    scale_size_continuous(range=c(2,0.5))

### Density plots for each part of speech for term coverage and click rate.
ggplot(data=HW2Q5_cleaned%>%
           pivot_longer(cols=c("Term.coverage","Click.rate"),
                        names_to="Variable",values_to="value"),
       mapping=aes(x=value,fill=Part.of.speech))+
    geom_density(alpha=0.5)+
    scale_x_log10()+
    facet_wrap(Variable~.,scales="free")
```

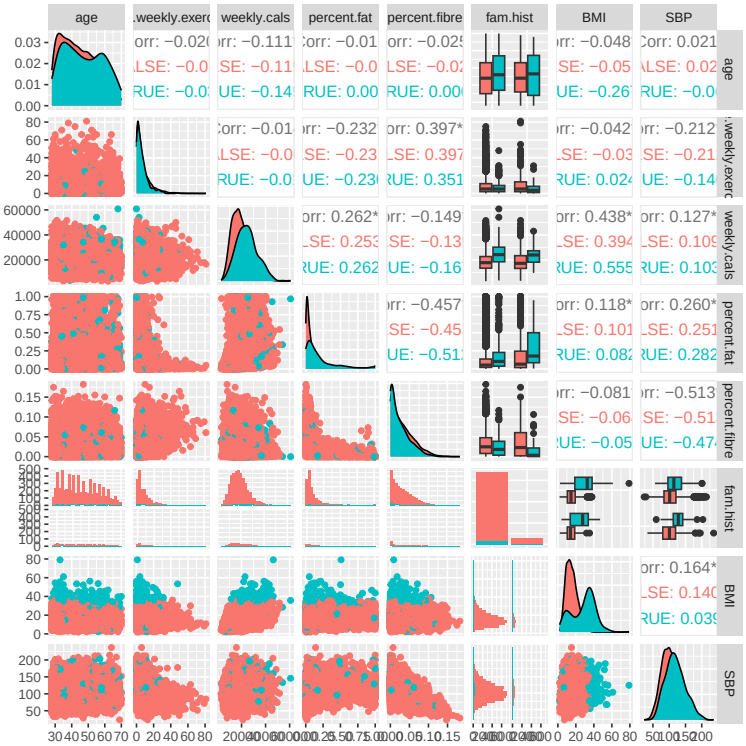
6. *The file* `HW2Q6.txt` *contains data from a study on the effect of exercise on the risk of heart disease in men. The variables included are*

Variable	*Meaning*
age	*The age of the patient*
`ave.weekly.exercise`	*The number of hours per week spent exercising.*
`weekly.cals`	*The number of calories consumed weekly.*
`percent.fat`	*The proportion of the patient's diet that consists of fats.*
`percent.fibre`	*The proportion of the patient's diet that consists of fibre.*
`fam.hist`	*Whether the patient has family history of heart disease.*
`BMI`	*The patient's BMI.*
`SBP`	*The patients systolic blood pressure.*
`heart.5.year`	*Whether the patient develops heart disease within the following 5 years.*

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.

We are not given information about the source of this data, so it is hard to be sure how reliable it might be. The lifestyle details are probably based on patient surveys, which are extremely unreliable, and could have bias. The medical measurements `BMI` and `SBP` are probably clinically measured and so are probably accurate at the time of recording. As these indices vary, there will be some error from using a single measurement, but it should be unbiased.

A quick examination of the data shows 5 missing values for family history. These do not show any particular pattern. None of them had heart attacks within the following 5 years, but given the frequency of heart attacks, that is not unusual. We therefore remove these points from the data. We also observe that there are 10 observations with `percent.fat` greater than 1. It is possible that these were input as percentages, rather than proportions. However, without verifying this, it is safer to remove these entries, which do not show any particular patterns for other predictors. There are also 11 additional entries where the sum of `percent.fat` and `percent.fibre` exceeds 1, which should not be possible. We also confirm that there are no duplicated records. After removing missing and incorrect values, we plot pairwise scatterplots coloured by whether the individual experienced a heart attack within 5 years.



From this, we see that `percent.fat`, `percent.fibre` and `ave.weekly.exercise` have skewed distributions, and might benefit from log-transformations. (Logistic transformation is an alternative possibility for `percent.fat` and `percent.fibre`.) After log-transforming these variables, we get the following pairwise scatterplots.

After the transformation, these variables are more normal, although log exercise is slightly bimodal, with a cluster of observations around 0. There are a few outliers in `weekly.cals`, `BMI` and `SBP`. There are also some outliers to the relationship between `SBP` and `log.exercise`. These outliers are not so extreme, so could be left in the dataset. We see that there is a negative association between `percent.fat` and `percent.fibre`, and positive associations between `BMI`, `SBP` and `weekly.cals`. These relations could be linear after log-transformation of the skewed predictors.

We have identified the following from our data analysis:

- Some of the predictors may be biased, due to data collection. Other predictors may be inaccurate.
- There are 5 missing values for `fam.hist`, which appear to be missing at random. I have removed these values.
- There are 10 impossible values for `percent.fat`, which are probably percentages instead of proportions, but I have removed them.
- There are no duplicated records.
- There is a negative association between `percent.fat` and `percent.fibre`
- There are positive associations between `BMI`, `SBP` and `weekly.cals`. These relations could be linear.

The data exploration was performed using the following code.

```
HW2Q6<-read.table("../HW2Q6.txt",stringsAsFactors=TRUE)
summary(HW2Q6)
HW2Q6[is.na(HW2Q6$fam.hist),]
HW2Q6[HW2Q6$percent.fat>1,]
which(duplicated(HW2Q6))


library(GGally)
library(dplyr)

HW2Q6_clean<-HW2Q6%>%filter(percent.fat+percent.fibre<=1,!is.na(fam.hist))


ggpairs(HW2Q6_clean%>%mutate(fam.hist=as.logical(fam.hist))%>%
        select(-c("heart.5.year")),
        mapping=aes(colour=HW2Q6_clean$heart.5.year))


ggpairs(HW2Q6_clean%>%mutate(log.exercise=log(ave.weekly.exercise),
                             log.fat=log(percent.fat),
                             log.fibre=log(percent.fibre),
                             fam.hist=as.logical(fam.hist))%>%
        select(-c("heart.5.year",
                  "ave.weekly.exercise",
                  "percent.fat",
                  "percent.fibre")),
        mapping=aes(colour=HW2Q6_clean$heart.5.year))
```