ACSC/STAT 3740, Predictive Analytics

WINTER 2024
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.]

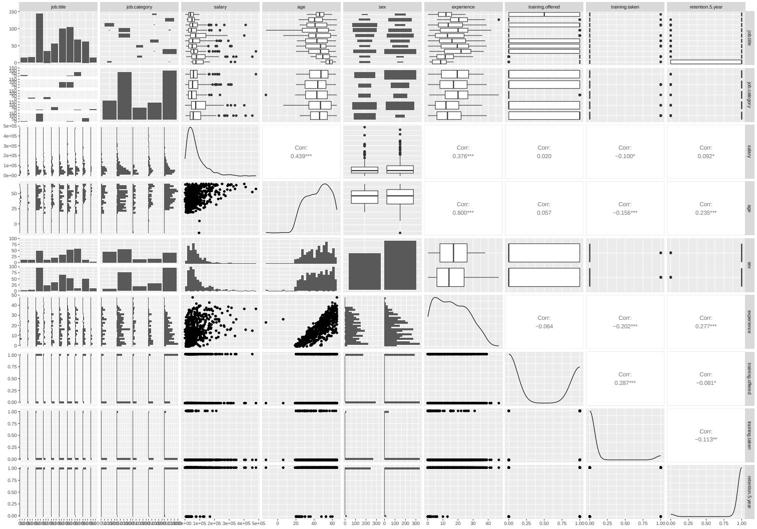
Standard Questions

1. *The file `HW2Q1.txt` contains the following data from a company's human resources department about employee retention*

Variable	*Meaning*
`job.title`	*The employee's job title.*
`job.category`	*The type of work.*
`salary`	*The employee's annual salary.*
`age`	*The employee's age.*
`sex`	*The employee's gender.*
`experience`	*The number of years of experience in the current job.*
`training.offered`	*Whether the employee was offered training courses.*
`training.taken`	*Whether the employee took the offered training courses*
`retention.5.year`	*Whether the employee is still working in the company after 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. You should take into account any concerns with data collection and processing.

We start with pairwise scatterplots.

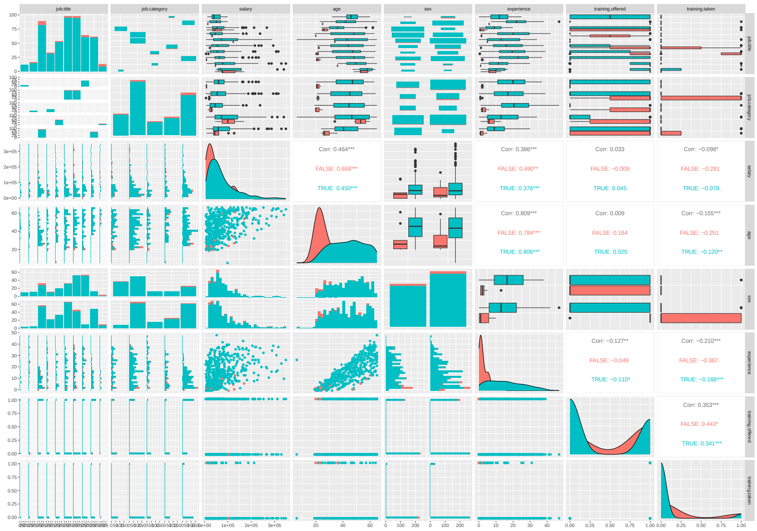


These highlight a few data issues. For most cases, a given job title falls entirely within one job category. However, there are a few cases where this does not happen. This might be a data error, or could be individuals with unusual jobs that do not fit the usual categorisation. We should investigate the effect of removing these individuals. There are some individuals with very low salaries. These can't represent annual salaries for full-time employees, but may represent special circumstances. There are also some outliers on the high end for salaries. This is possible, as salaries can have very heavy-tailed distributions. However, some of these are in job titles that do not usually have such large salaries, and do not have such large experience. There are two clear outliers in age, with negative values which are clearly wrong.

Another issue we notice is that there are 4 individuals with `training.taken=TRUE` and `training.offered=FALSE`. From the definitions given, this should be impossible. It is unclear what the mistake is, so I will remove these observations.

We remove these outliers. The filtering also removes the NA values in salary and age, which could influence the results. There are 51 missing values for salary and 3 missing values for age. The missing values for salary are all engineers with 20 or more years of experience. They were all retained. The missing values for age are all male, though this might be coincidence. Two pairs of the entries with missing salary are duplicates, but this is most likely a coincidence, since most of the non-missing variables are factors or logical. Given the non-random missing pattern for salary, there is a danger that the removal could bias the results. We should consider using other methods to deal with these missing values if the correct values cannot be obtained.

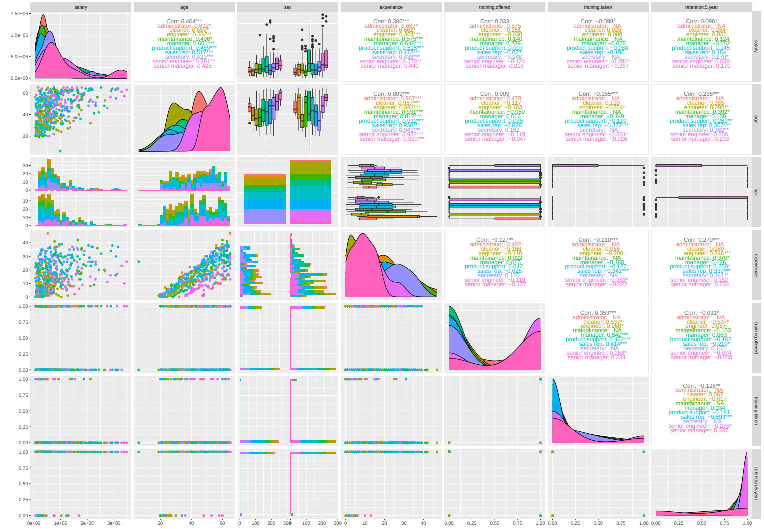
We next examine pairwise scatterplots coloured by 5-year retention



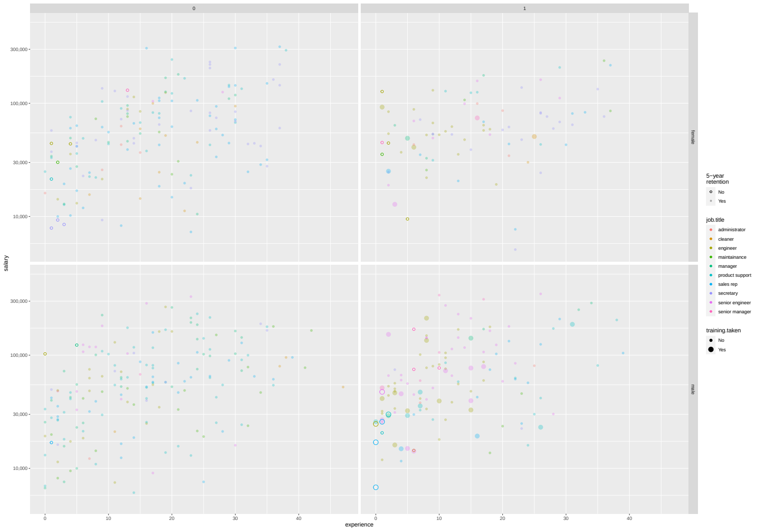
We see that there is a strong linear relation between age and experience. Retention rates are fairly high, but much lower for employees without much experience. Salary is heavy-tailed and should be log-transformed. Salary is correlated with experience and job title.

Given the linear relation between age and experience, we might consider adding a feature for the difference between age and experience. However, in this case, it is not particularly natural.

Colouring the plots by job title shows that many things are affected by job title. I used the default colour scale for this plot, but it may be worth designing a colour scale so that jobs in the same job category have similar colours.



We see that for any given job title, age and experience are extremely highly correlated, so we may not need both as predictors. We can try to show all predictors on a single plot to look for additional patterns.



Since retention is high, the number of employees who left is relatively small, so it is difficult to judge patterns. However, we see there are clear differences between male and female employees. Male employees were more likely to be offered training, and more likely to take it if offered. They were also more likely to leave after taking training. Employees with little experience were much more likely to leave. The effect of salary on retention is unclear.

In summary, the data exploration found the following:

- There are several cases where a job title is placed in an unusual job category.
- There are a large number of missing values for salary. These are all from engineers with over 20 years of experience, so removing these observations could give biased results.
- There are a few missing values for age. These appear to be random, so can be removed.
- There are several individuals who were not offered training, but took the training. This seems like a mistake, so we remove these individuals from the data.
- There are some individuals with negative age — clearly a mistake.
- There are some outliers in salary. At the high end, these are possible, but unlikely. At the low end, these seem infeasible for full-time employees, and are either mistakes or special circumstances that should be handled differently. We should therefore remove these values.

- Age and experience are very strongly correlated, particularly when divided by job title.
- Salary has a heavy-tailed distribution, so a transformation might be appropriate.
- Experience is a very important predictor of retention, with experienced employees much less likely to leave. Sex is also an important predictor, and seems to have interaction with training offered and training taken, so we may need to include some interaction terms in our model.

The code used for this exploration is the following:

```
HW2Q1<-read.table("HW2Q1.txt")
library(GGally)
ggpairs(HW2Q1)

summary(HW2Q1)
table(HW2Q1$job.category,HW2Q1$job.title)

table(HW2Q1$training.offered,HW2Q1$training.taken)

which(duplicated(HW2Q1))
HW2Q1[c(581,609),]
HW2Q1%>%filter(is.na(salary))
HW2Q1%>%filter(is.na(age))

library(dplyr)
HW2Q1_clean<-HW2Q1%>%filter(age>0)%>%
    filter(salary<375000)%>%filter(salary>5000)%>%
    filter((job.title!="administrator")|(job.category=="admin"))%>%
    filter((job.title!="engineer")|(job.category=="technical"))%>%
    filter((job.title!="sales rep")|(job.category=="customer"))%>%
    filter((job.title!="secretary")|(job.category=="admin"))%>%
    filter(training.offered|!training.taken)

ggpairs(HW2Q1_clean%>%select(-c("retention.5.year")),
        mapping=aes(colour=(HW2Q1_clean$retention.5.year==1)))


ggpairs(HW2Q1_clean%>%select(-c("job.title","job.category")),
        mapping=aes(colour=(HW2Q1_clean$job.title)))


ggplot(HW2Q1_clean,
       mapping=aes(x=experience,
                   y=salary,
                   colour=job.title,
                   alpha=as.logical(retention.5.year),shape=as.logical(retention.5.year),
                   size=training.taken))+
    geom_point()+
    facet_grid(sex~training.offered)+
    scale_y_log10(labels=scales::comma,limit=c(5000,500000))+
    scale_shape_manual(values=c(1,16),
                       name="5-year\nretention",
                       labels=c("No","Yes"))+
    scale_alpha_manual(values=c(1,0.3),
                       name="5-year\nretention",
                       labels=c("No","Yes"))+
    scale_size(breaks=c(0,1),range=c(2,4),labels=c("No","Yes"))
```

2. *The file* `HW2Q2.txt` *contains the following data about school performances in standardised tests for Grade 8:*

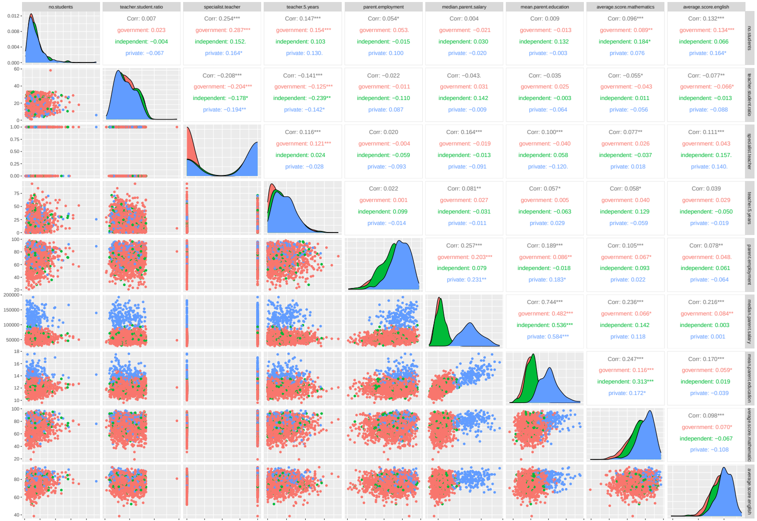
Variable	*Meaning*
no.students	*The number of students in Grade 8 attending the school.*
teacher.student.ratio	*The average number of students per teacher in a class at the school.*
funding	*The schools source of funding — government, independent or private.*
specialist.teacher	*Whether the school employs teachers with specialist knowledge for each subject.*
teacher.5.years	*The percentage of teachers at the school with at least 5 years of experience.*
parent.employment	*The percentage of parents of children at the school who are employed.*
median.parent.salary	*The median salary of parents of children at the school*
mean.parent.education	*The average number of years of full-time education of parents of children at the school.*
average.score.mathematics	*The average score of children in Grade 8 at the school on the standardised mathematics test.*
average.score.english	*The average score of children in Grade 8 at the school on the standardised English test.*

The test results were published by the examination board. Information on schools was provided by the schools. Information about parents of children was taken from surveys conducted by the schools.

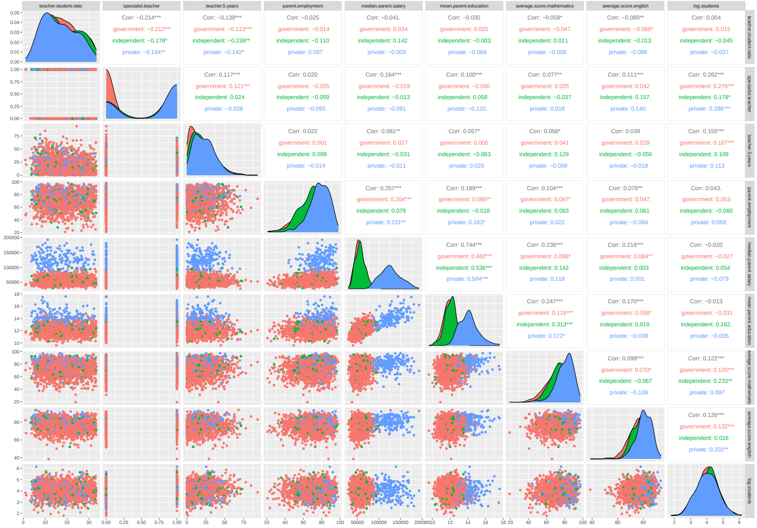
Perform data exploration on this data set, and summarise (with tables and plots to support where appropriate) your initial conclusions about data issues and appropriate models. You should take into account any concerns with data collection and processing.

The data set brings together data from a number of sources, which could cause issues. The calculation of some variables may not be completely standard. For example, for teacher-student ratio, what classes are included? Could a school manipulate this by holding very small classes for one hour each year, or is it time-weighted? Does this calculation differ between schools? Survey data always induces sampling bias. Furthermore, this sampling bias might be different for different schools. For example the number of parents unwilling to answer questions about their salary might be different at government-funded and private schools.

We start by plotting pairwise scatterplots. We use colour to indicate the school funding.



We see a clear outlier in teacher.student.ratio. This should probably be removed from the data. There is also an outlier in no.students, but that is heavy-tailed, so we should consider transforming that variable, and decide whether that school is an outlier after transformation.



After transformation, the school is not an outlier. There are no missing values in the dataset. Checking for duplicates, we see that the following rows are duplicates: 678 and 679, 728 and 729, 890 and 891, and 1018 and 1019. Given the number of numerical variables, and the fact that the duplicates are consecutive, it is too implausible that these could be genuine records. They must surely be data errors, so we should remove the duplicates.

We see that the parent variables are very different for private schools, as might be expected. Parent salary and education are strongly correlated. The correlation between the scores in the two subjects is relatively low. The best predictors for average mathematics score appear to be parent education, school funding and number of students; while the best predictors for English score appear to be parent salary, school funding and number of students. After log-transforming the number of students the relations between predictors seem to be mostly linear, suggesting a linear model may be appropriate. Furthermore, the linear models may be similar for different school types, though in some cases there are fairly large differences between correlation coefficients for different school types. This may be because most schools are government-funded, meaning that the coefficients for private and independent schools are more variable. However, it may be worth including interaction terms between funding source and other predictors.

The conclusions from the data exploration are:

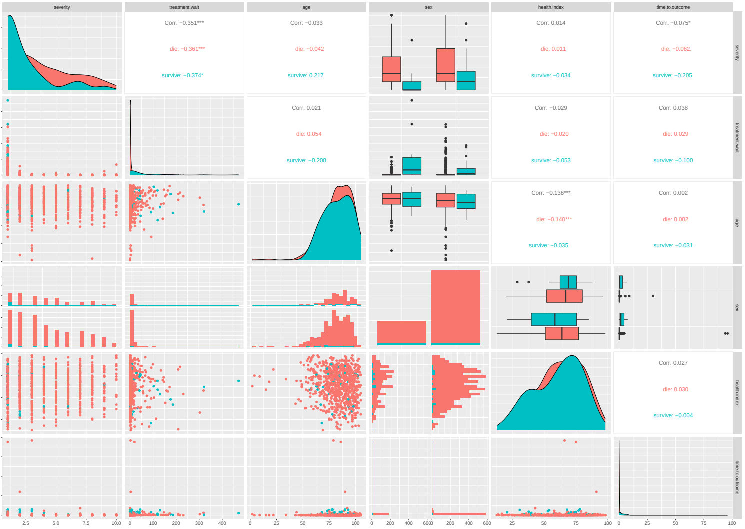
- There are some potential issues with combining the data sources. We should check how teacher-student ratio is defined for each school to ensure the values are comparable. The parent variables are collected from surveys, so may be subject to sampling bias.
- There are duplicated records. These are almost certainly data entry problems and should be removed.
- There is an outlier in teacher-student ratio, which should be removed.
- Number of students is heavy-tailed and should probably be log-transformed.
- For mathematics score, the best single-variable linear predictors are parent education, school funding and log number of students.
- For English score, the best single-variable linear predictors are parent salary, school funding and log number of students.
- There is relatively low correlation between average mathematics score and average English score.
- Most relations appear to be approximately linear.
- There may be interactions between school funding and other predictors.

3. *The file* `HW2Q3.txt` *contains the following data from an experiment about survival times under a certain treatment for a disease.*

Variable	*Meaning*
severity	*The severity of disease symptoms prior to treatment 1=mild, 10=very severe.*
treatment.wait	*The number of days the patient needs to wait before being treated.*
age	*The patient's age at first diagnosis.*
sex	*The patient's sex.*
health.index	*An index assessing the patient's overall health — 100=perfect health, 0=dead*
outcome	*Whether the patient dies before the end of the study period.*
time.to.outcome	*The time in years before the patient dies or the end of the study, whichever happens first.*

Severity is assessed by the treating physician at first consultation. Treatment wait is calculated as the difference between the date of the first consultation and the start of treatment, from the hospital database. Age and sex are from the patient's medical records. Health index is based on physician assessment and a questionaire filled out by the patient. Outcome and time to outcome are from the hospital records.

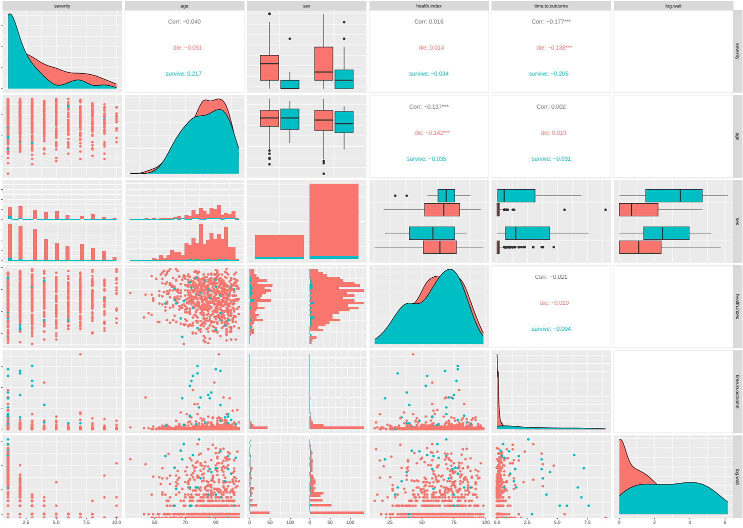
Most of the data are from hospital records, and should be fairly reliable. Health index and severity are based on physician assessment, so they may be slightly subjective, and there may be variation between physicians. This should not greatly affect the data analysis, but it is possible that the interpretation is changed, since some physicians may be more skilled than others, and some phyisician's may give biased assessment of these variables. A quick check shows that there are no missing values in the data. We start with pairwise scatterplots coloured by outcome.



We immediately see three outliers in time to outcome, with values over 30 — longer than the length of the study. It is possibly that these are in days.

We remove these outliers. We see that severity has a skewed distribution, with most patients diagnosed before symptoms become too severe. We see that treatment wait drops very quickly as severity increases. There are however several outliers with severity 10 but long treatment waits. These may be errors in the data, or may be because such severe symptoms cause difficulties starting treatment. We should examine these outliers carefully to see whether they appear to be anomalous in other ways. There are also some outliers in age, with some ages less than 10. Usually paediatric cases would be removed from the dataset, so these cases may be errors, and we will remove them in either case. To be thorough, we remove all ages less than 25.

After removing the outliers in time to outcome and age and log-transforming treatment wait, we replot the pairwise scatterplots.

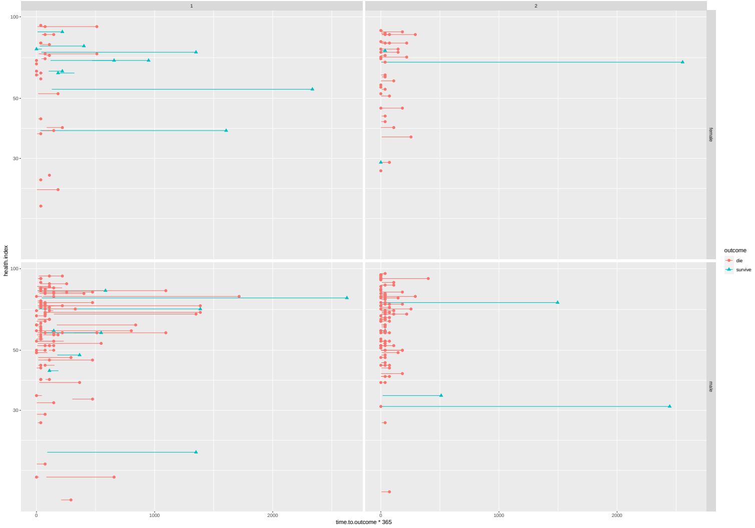


There is still one outlier in age. It is probably not too influential, but we might consider removing it. We see there is a very clear relation between severity and log-transformed wait times, with a few outliers with wait times much longer than usual for the severity. Here is a table of these outliers (selected by the formula $\log(\texttt{treatment.wait}) * (\texttt{severity} - 0.5) > 5$)

Number	Severity	Treatment wait	Age	Sex	Health index	Outcome	Time to Outcome
38	8	3	79	male	59	die	0.7
409	9	4	85	male	61	die	0.0
439	10	66	91	female	96	die	0.2
558	5	14	99	male	51	die	0.3
608	10	4	97	male	71	die	1.0
690	10	2	83	male	53	die	0.1
713	9	24	60	male	66	die	0.2
801	2	56	90	male	63	die	0.2

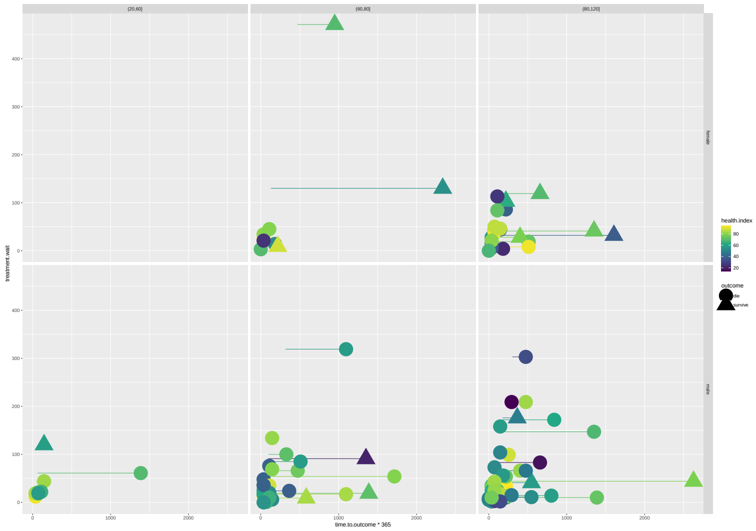
Some are only a few days and are probably reasonable. Data points 439, 713 and 801 are a little strange, so we remove these outliers. We also see that the disease affects men more than women, and a higher proportion of men died during the study period. We see that higher severity is associated with a higher chance of dying. Shorter wait times also appear to be associated with higher chance of dying, possibly because they are more severe cases. Other variables do not appear to be strongly associated with chance of dying.

Since we have two times (to treatment and to outcome), it may be appropriate to use a pointrange plot to show both.



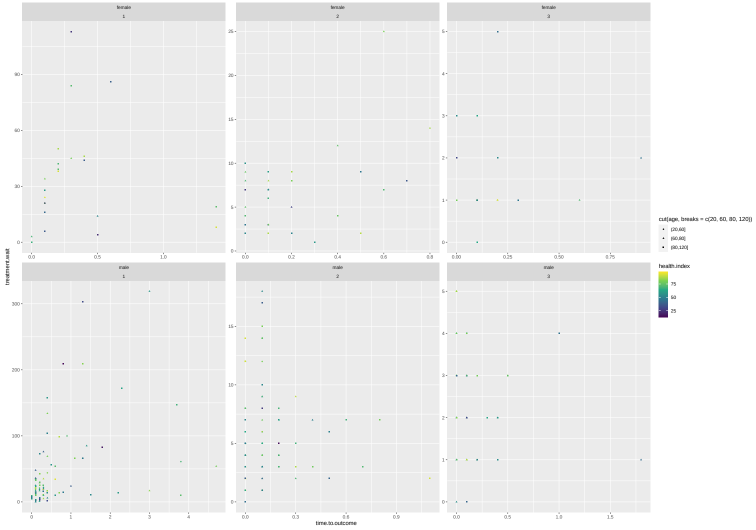
We note that there are some cases where the time to outcome is less than the treatment wait. The time to outcome is in years and rounded to one decimal place, so there could be rounding effects, but we see that in some cases, the time to outcome is hundreds of days less than time to treatment. This may indicate that treatment was planned but not performed, or may be an error. Since the focus of the study is the effect of the treatment, it makes sense to remove patients who left or died before receiving treatment. Because of rounding errors, it is difficult to decide on an exact cut-off.

Since survival rates are so low for higher severity, we restrict to the case severity=1. This plot may be clearer if we arrange points by wait time. If we increase the size of points, we can use colour to indicate health index and use one facet to group ages.



We see that even restricted to severity 1 cases, longer wait times are associated with higher survival chance.

An alternative approach would be to ignore the individuals who do not die, and plot time to death against time to treatment for all individuals. We remove higher severities as these do not have enough data points.



This shows a similar pattern with longer wait times associated with longer survival times, even when considering factors such as age and health index.

There are specialised models for this type of data — the individuals who leave the study are censored, and need to be treated appropriately.

A summary of the conclusions of the data exploration are the following:

- There are no missing values.
- There are 3 outliers in time to treatment, and a small number of outliers in age, that are either mistakes, or paediatric cases that should be handled separately. These outliers should be removed.
- There are a number of patients with high severity, but long wait times. We perform further checking to determine whether they are anomolous in other ways. There is nothing very obvious, and some are only a few days. We remove three outliers that have long wait times.
- The disease affects more men than women, and men appear to have higher risk of dying. Higher severity is also associated with higher chance of dying. Even for patients with severity 1, longer wait times are associated with lower chance of dying, and with longer time until death.
- There are a number of patients with wait times longer than time to outcome. We removed these from the data.
- It would be natural to use survival models to model this data set.

4. *A credit card company is improving its fraud detection models. It collects the following data for all purchases made:*

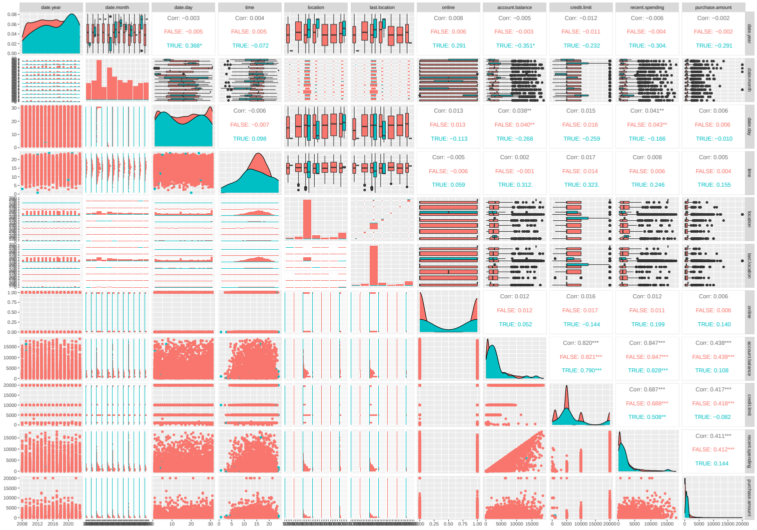
Variable	*meaning*
date	*The date of the transaction.*
time	*The time of the transaction.*
location	*The location of the transaction.*
last.location	*The location of the previous transaction from this credit card.*
online	*Whether the transaction was online.*
account.balance	*The amount owing on the card before the transaction.*
credit.limit	*The credit limit of the card.*
recent.spending	*The amount spent on the card in the previous week.*
purchase.amount	*The cost of the purchase.*
fraudulent	*Whether the purchase was fraudulent.*

Most of the data are automatically processed at time of purchase. Fraudulent is based on customer reports and manual reviews. The data are in the file `HW2Q4.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.

The automatically collected data should be unbiased. Any processing errors are likely to be random. For fraudulent transactions, it is possible that some fraudulent purchases were unnoticed by the customers, or that the manual reviews were inconclusive in some cases, which could lead to fraudulent cases not labelled as fraudulent. Non-fraudulent purchases labelled as fraudulent seems less likely, but not impossible.

A quick summary of the data shows that there are no missing values. We also see that the data are very unbalanced, with only 32 fraudulent uses out of almost 6000 transactions. We start with pairwise scatterplots, using colour to indicate the fraudulent transactions.

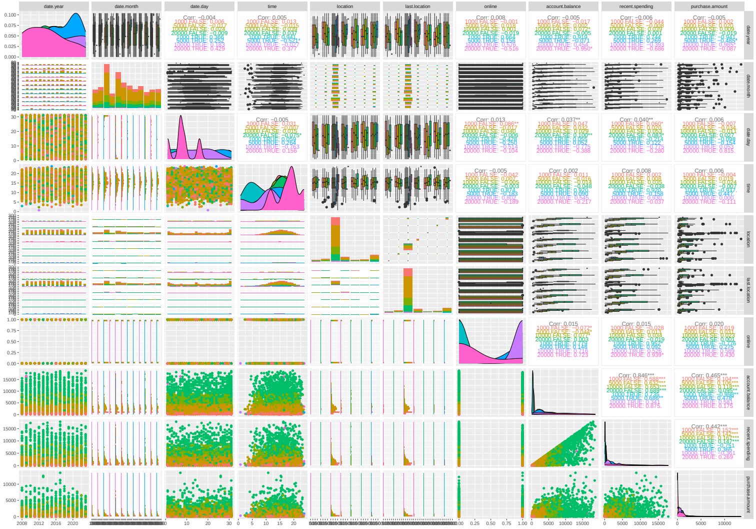


We notice that there are a number of transactions where account balance, recent spending and purchase amount all exceed credit limit, which seems implausible. Furthermore, most of these transactions also have credit limit not a round number, and purchase amount a round number, suggesting that purchase amount and credit limit have been swapped. As these cases are not fraudulent, and there are not many of them, we remove these cases, rather than attempt to fix them.

We see that the rate of fraudulent claims has been increasing. Payments are not evenly distributed over the year, being much more frequent in December and January. Most payments in December are towards the end of the month, and most in January are towards the beginning, but in other months, the payments are uniformly distributed. Most payments are made between 5:00 in the morning and 10:00 in the evening, with payments outside these hours more likely to be fraudulent. Most payments occur in Canada, and most payments occur in the same location as the last payment. About 46% of purchases are made online, with similar proportions of fraudulent payments online and in-person. We see that account balance is always less than credit limit, and recent spending is less than account balance. Purchase amounts have a heavy-tailed distribution, and may benefit from log-transformation. After removing the problematic data, there are only two cases where purchase amount exceeds credit limit. In both cases, credit limit is $1,000 and the transactions were fraudulent.

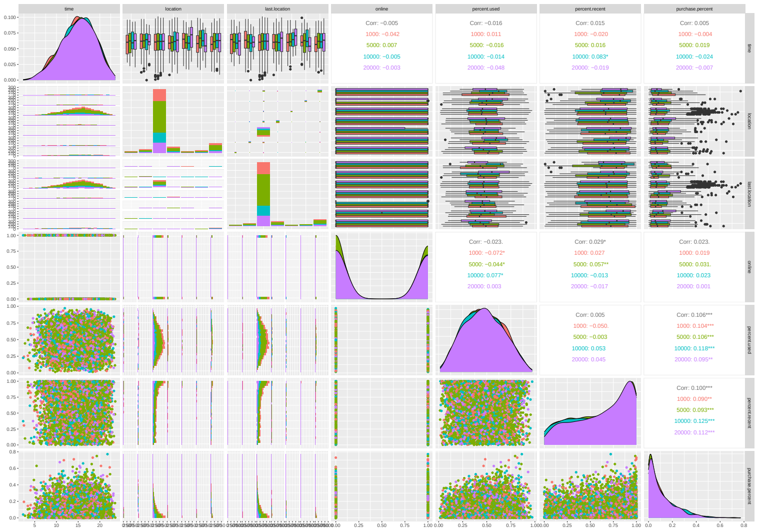
Higher purchase amounts are more likely to be fraudulent.

Given how related all the variables are to credit limit, it may be appropriate to calculate balance, recent spending and purchase amount as a percentage of credit limit, or recent spending as a percentage of account balance. Since credit limit only has a limited number of values, we can colour the pairwise scatterplots by the interaction of credit limit and fraud status.



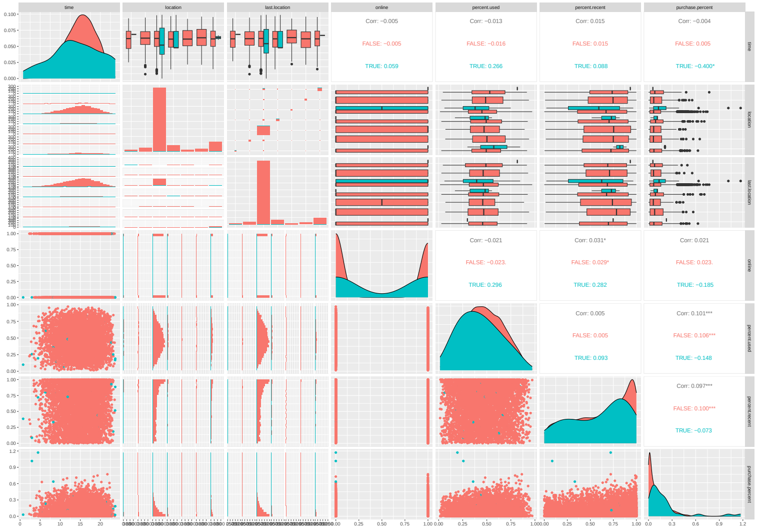
This plot is a little difficult to read. The correlation between account balance, recent spending and purchase amount is reduced when considering a single credit limit, but is still significant.

We plot the new features. To simplify the figure, we exclude the fraudulent transactions and colour by credit limit.



From this plot, we see that the transformed features are much more stable over credit limits, and there is still a low correlation between purchase percent and both percent used and percent recent. Purchase percent still has a long-tailed distribution, so may benefit from log transformation.

We next plot the same transformed features, but coloured by fraud status. We also set alpha to 0.5 to make it easier to see the overlapping histograms.



We see that several fraudulent transactions appear as outliers in time or purchase percent. [The density plots for online look different, but this is an artefact of bandwidth selection due to low sample size for fraudulent transactions.]

We have identified the following conclusions:

- The data are very unbalanced with only 32 fraudulent transactions in the data set.
- There are some problematic records with purchase amount and credit limit seeming to be switched.
- The rate of fraudulent claims has been increasing over time.
- Account balance, recent spending and purchase amount are all strongly correlated with credit limit. Creating new features by expressing these as percentages of credit limit or other amounts may be helpful. After doing this, these features have similar distributions for all credit limits.
- Payments are not evenly distributed over the year, being much more frequent in December and January. Most payments in December are towards the end of the month, and most in January are towards the beginning, but in other months, the payments are uniformly distributed.
- Most payments are made between 5:00 in the morning and 10:00 in the evening, with payments outside these hours more likely to be fraudulent.
- Most payments occur in Canada, and most payments occur in the same location as the last payment.
- Purchase amount divided by credit limit still has a long-tailed distribution, so may benefit from log transformation.
- There are some outliers in time and purchase percent, which are fraudulent.

5. *A scientist is studying the effect of climate on road maintainance costs. She collects the following data from a large number of cities*

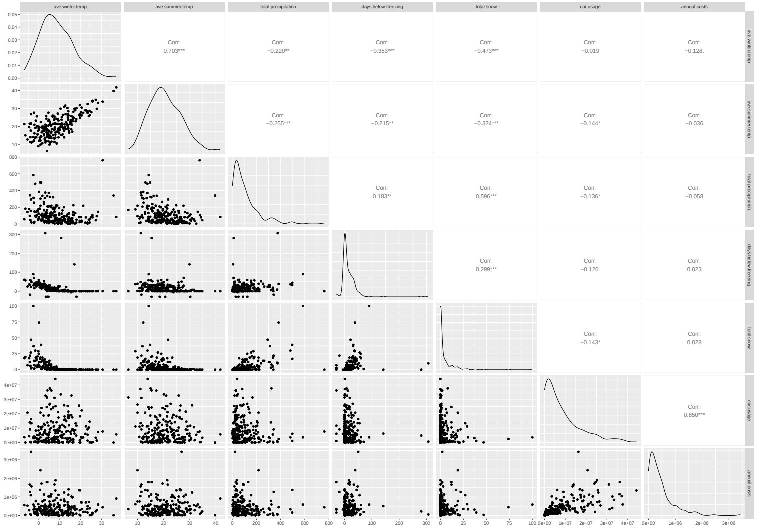
Variable name	*Meaning*
ave.winter.temp	*The average daily maximum temperature for the months January–March.*
ave.summer.temp	*The average daily maximum temperature for the months June–August.*
total.precipitation	*The total yearly rainfall (mm)*
days.below.freezing	*The number of days during the year where the maximum temperature is below* $0°C$.
total.snow	*The total annual snowfall (cm)*
car.usage	*The total distance driven annually by all inhabitants of the city. (km)*
annual.costs	*The total annual costs for road maintainance ($).*

The data are in the file `HW2Q5.txt`. *The weather data are from historical weather records at nearby weather stations. Car usage is estimated from surveys. Annual costs are from local government reports.*

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.

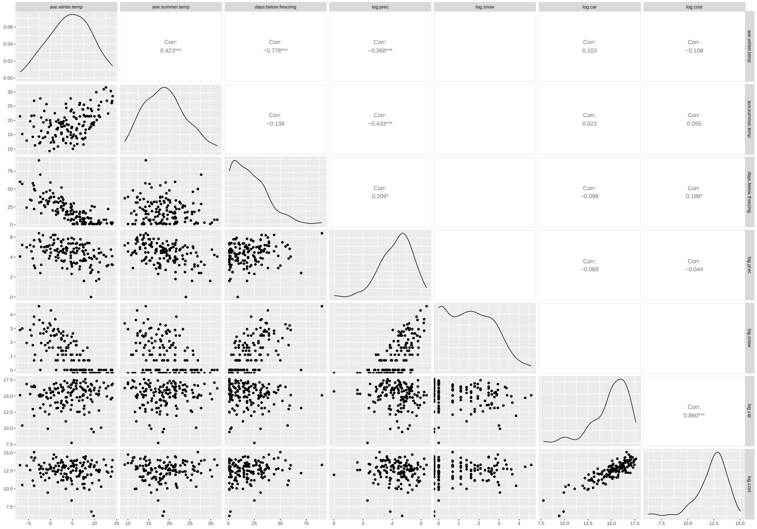
Recent weather records are usually quite reliable, and would not be expected to be subject to bias. However, in some cases, the weather stations are not as close to the city, which could result in some measurements being inaccurate. Surveys can be very unreliable. There is not an obvious direction for possible bias, but sampling bias could significantly affect the estimated car usage. The reliability of local government reports can vary. In some places, the reports may be deliberately false. In other places, different accounting methods could result in the values not being perfectly comparable — expenses classified as road maintainance by one council may be differently classified by another.

There are no missing values and no duplicate values in the dataset. We start by plotting pairwise scatterplots.



We see a large number of outliers. There are some cases with a negative number of days below zero, which are clearly errors. We remove these points. There are also three cities with over 100 days below freezing. Given that these cities have average winter temperatures above zero, this is probably a mistake, and we also remove these points. There are two cities with average summer and winter temperatures of about $40°C$, which is not impossible, but quite extreme. We should check how much these observations influence the fitted model. Total precipitation and total snow are very heavy-tailed, and should probably be log-transformed. There are potential outliers for these variables, but these should be reassessed after the log-transformation. Car usage is also heavy-tailed and should be log-transformed.

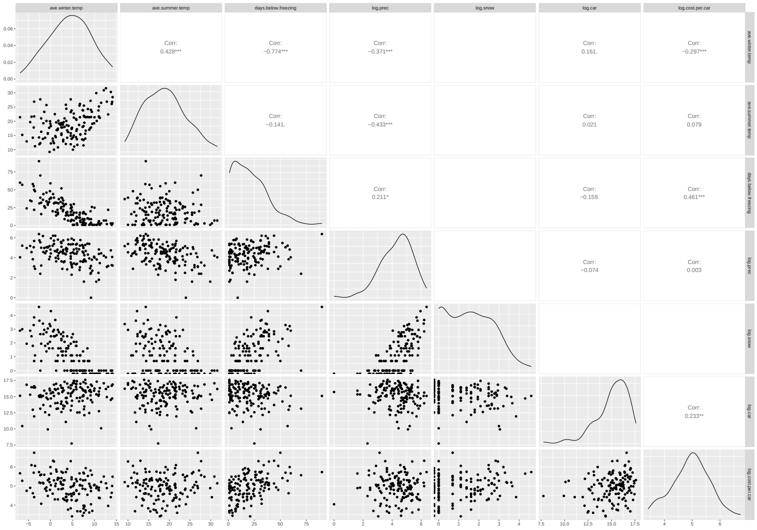
After removing the outliers and log-transforming the heavy-tailed variables, we replot pairwise scatterplots.



After log transformation, the large values for total precipitation and total snow are no longer outliers. There is a very strong linear relation between log-transformed car usage and log-transformed annual costs. The slope of this is close to 0.5, so it might be appropriate to consider the transformed variable cost/$\sqrt{\text{car usage}}$. After removing outliers, days below freezing is still heavy-tailed, so may benefit from log transformation.

Other variables are not significantly correlated with log-transformed annual cost, with the possible exception of days below freezing. This could be because they are not correlated with car usage. There is quite strong correlation between average winter temperature and days below freezing. It may be appropriate to remove one from the model.

When we plot the feature cost/$\sqrt{\text{car usage}}$, we see there are two outliers in this feature, where the annual costs are very low. After removing these outliers and replotting pairwise scatterplots:



We see that log days below freezing and average winter temperature are strongly correlated with the transformed variable.

Thus, the conclusions to our data exploration are:

- There are no missing values or duplicates.
- There are some impossible values for days below zero, and some implausible values, which should all be removed.
- There are two slight outliers in temperature, that we do not remove, but we should monitor the influence of these outliers.
- Total precipitation, total snow, days below freezing, car usage and annual costs are heavy-tailed and should be log-transformed.
- Car usage is the best predictor of annual costs, and after log-transformation is very strongly linearly related with annual costs.
- After removing the effect of car usage, there is an association between average winter temperature or log-transformed days below freezing and log transformed annual cost. This association is fairly linear
- Average winter temperature and days below freezing are strongly negatively correlated.