LEAD example

Niall Anderson

2022

This is an example of a R Markdown document you could use to carry out an analysis. If using this template to write a report of your findings, you might want to structure it a little differently. For example, you could choose to hide the R code for the preamble, data check and data preparation and some of the Exploratory Data Analysis, to keep only sections such as:

- Introduction
- Materials and Methods
- Results – This could have the R outputs displayed but the R code hidden
- Discussion
- Conclusion

Or any other structure!

Preamble R code

```
# Load the packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(finalfit)

# Import the dataset
Lead <- read_csv("LeadAb.csv")
```

```
## Rows: 501 Columns: 37
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (37): idno, bloodpb, abscore, school, ageint, sex, moveschl, classyr, ti...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data checking

It can be useful to inspect the data by eye, you can often spot rogue values, data entry issues, etc.

```
# View the dataset in RStudio
view(Lead)
# Get summary statistics for all variables
summary(Lead)
```

```
##       idno           bloodpb         abscore        school
##  Min.   :  33.0   Min.   : 33.0   Min.   : 68   Min.   : 3.00
##  1st Qu.: 376.0   1st Qu.: 89.0   1st Qu.:103   1st Qu.: 8.00
##  Median : 756.0   Median :117.0   Median :113   Median :12.00
##  Mean   : 741.2   Mean   :123.6   Mean   :112   Mean   :11.58
##  3rd Qu.:1108.0   3rd Qu.:148.0   3rd Qu.:121   3rd Qu.:15.00
##  Max.   :1502.0   Max.   :340.0   Max.   :146   Max.   :20.00
##      ageint            sex           moveschl         classyr
##  Min.   : 81.00   Min.   :1.000   Min.   :0.0000   Min.   :3.000
##  1st Qu.: 90.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:3.000
##  Median : 96.00   Median :1.000   Median :0.0000   Median :4.000
##  Mean   : 96.19   Mean   :1.479   Mean   :0.0499   Mean   :3.503
##  3rd Qu.:102.00   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:4.000
##  Max.   :113.00   Max.   :2.000   Max.   :1.0000   Max.   :4.000
##     timeday          handed        famhist           fsoc           msoc
##  Min.   :1.000   Min.   :1.00   Min.   :0.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.000   Median :1.00   Median :0.000   Median :2.000   Median :2.000
##  Mean   :1.413   Mean   :1.12   Mean   :0.492   Mean   :2.116   Mean   :1.836
##  3rd Qu.:2.000   3rd Qu.:1.00   3rd Qu.:0.000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.00   Max.   :5.000   Max.   :4.000   Max.   :4.000
##     mqualif         fqualif          unempl         workmum
##  Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :2.000   Median :2.000   Median :1.0000
##  Mean   :2.405   Mean   :2.641   Mean   :1.954   Mean   :0.7645
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :5.000   Max.   :5.000   Max.   :2.000   Max.   :2.0000
##     parhlth          parment           totcigs         carphone
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.00   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.00   1st Qu.:1.000
##  Median :0.0000   Median : 0.0000   Median : 1.00   Median :2.000
##  Mean   :0.3852   Mean   : 0.4391   Mean   :11.77   Mean   :1.593
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:20.00   3rd Qu.:2.000
##  Max.   :6.0000   Max.   :10.0000   Max.   :80.00   Max.   :2.000
```

```
##     consumer        occuprat       famsize         birthord
##  Min.   :0.000   Min.   :0.29   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:0.80   1st Qu.:2.000   1st Qu.:1.000
##  Median :2.000   Median :1.00   Median :2.000   Median :2.000
##  Mean   :2.359   Mean   :1.05   Mean   :2.128   Mean   :1.675
##  3rd Qu.:3.000   3rd Qu.:1.33   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :3.00   Max.   :3.000   Max.   :3.000
##      gestat           brthwt           brthsco          medhist
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.0978   Mean   :0.05589   Mean   :0.5289   Mean   :0.2036
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :2.0000   Max.   :1.00000   Max.   :3.0000   Max.   :1.0000
##     stheight          offschl          childint        parchcom
##  Min.   :-2.6400   Min.   : 0.000   Min.   :0.000   Min.   :1.000
##  1st Qu.:-0.3900   1st Qu.: 3.000   1st Qu.:2.500   1st Qu.:5.000
##  Median : 0.1900   Median : 6.000   Median :3.250   Median :6.000
##  Mean   : 0.2103   Mean   : 8.884   Mean   :3.207   Mean   :6.106
##  3rd Qu.: 0.7500   3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:7.000
##  Max.   : 3.2900   Max.   :65.000   Max.   :6.500   Max.   :8.000
##     parpart          parschl           pvoc            pmat
##  Min.   :1.000   Min.   :0.000   Min.   : 4.00   Min.   :13.0
##  1st Qu.:4.500   1st Qu.:2.000   1st Qu.:43.00   1st Qu.:38.0
##  Median :5.500   Median :3.000   Median :52.00   Median :45.0
##  Mean   :5.246   Mean   :3.214   Mean   :52.48   Mean   :43.3
##  3rd Qu.:6.500   3rd Qu.:4.000   3rd Qu.:61.00   3rd Qu.:50.0
##  Max.   :6.500   Max.   :6.000   Max.   :84.00   Max.   :60.0
```

It's a useful quick summary of a data set, it will point out NA frequency etc. We can check whether we get the expected range for categorical variables (e.g. the Codebook says 1,2 but seeing 3, 4, 99 as well, etc.).

Data preparation

We will convert the categorical variables to factors. As levels for these variables are listed in numerical order (0, 1, 2, etc.), we can use the Codebook to replace those by more meaningful factor levels. This is useful to check level frequencies, etc. For this step, we will create a new working tibble, named Leadw, so that the original data is untouched. This allows to go back to the Lead object if there is a mistake in Leadw.

```
# Convert categorical variables to factors
# Create a working data frame (adding a suffix "w")
Leadw<-Lead %>%
  mutate(school=factor(school),
         sex=factor(sex,labels = c("Male","Female")),
         moveschl=factor(moveschl,labels = c("No","Yes")),
         classyr=factor(classyr,labels = c("Primary 3","Primary 4")),
         timeday=factor(timeday,labels = c("AM","PM")),
         handed=factor(handed,labels = c("Right","Left")),
         msoc=factor(msoc),
         fsoc=factor(fsoc),
         mqualif=factor(mqualif,labels = c("None","Apprentice","School","Higher School","FE","Degree")),
         fqualif=factor(fqualif,labels = c("None","Apprentice","School","Higher School","FE","Degree"))
```

```
        unempl=factor(unempl,labels = c("Yes","No")),
        workmum=factor(workmum,labels = c("No","PT","FT")),
        carphone=factor(carphone,labels = c("Neither","Only 1","Both")),
        gestat=factor(gestat,labels = c("38 Weeks+","34-37 Weeks","<34 Weeks")),
        brthwt=factor(brthwt,labels = c("No","Yes")),
        medhist=factor(medhist,labels = c("No","Yes"))
        )
# Check the summary statistics of the transformed dataset
summary(Leadw)
```

```
##       idno           bloodpb         abscore        school        ageint
##  Min.   :  33.0   Min.   : 33.0   Min.   : 68   13     : 54   Min.   : 81.00
##  1st Qu.: 376.0   1st Qu.: 89.0   1st Qu.:103   15     : 48   1st Qu.: 90.00
##  Median : 756.0   Median :117.0   Median :113   11     : 44   Median : 96.00
##  Mean   : 741.2   Mean   :123.6   Mean   :112   17     : 35   Mean   : 96.19
##  3rd Qu.:1108.0   3rd Qu.:148.0   3rd Qu.:121   9      : 34   3rd Qu.:102.00
##  Max.   :1502.0   Max.   :340.0   Max.   :146   7      : 28   Max.   :113.00
##                                                 (Other):258
##      sex       moveschl       classyr    timeday    handed       famhist
##  Male  :261   No :476   Primary 3:249   AM:294   Right:441   Min.   :0.000
##  Female:240   Yes: 25   Primary 4:252   PM:207   Left : 60   1st Qu.:0.000
##                                                             Median :0.000
##                                                             Mean   :0.492
##                                                             3rd Qu.:0.000
##                                                             Max.   :5.000
##
##  fsoc    msoc             mqualif             fqualif    unempl    workmum
##  1:217   1:198   None         : 88   None         : 87   Yes: 23   No:185
##  2: 56   2:227   Apprentice   :111   Apprentice   : 92   No :478   PT:249
##  3:181   3: 36   School       : 85   School       : 91             FT: 67
##  4: 47   4: 40   Higher School: 41   Higher School: 35
##                  FE           : 79   FE           : 36
##                  Degree       : 97   Degree       :160
##
##     parhlth          parment           totcigs         carphone
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.00   Neither: 28
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.00   Only 1 :148
##  Median :0.0000   Median : 0.0000   Median : 1.00   Both   :325
##  Mean   :0.3852   Mean   : 0.4391   Mean   :11.77
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:20.00
##  Max.   :6.0000   Max.   :10.0000   Max.   :80.00
##
##     consumer        occuprat       famsize         birthord
##  Min.   :0.000   Min.   :0.29   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:0.80   1st Qu.:2.000   1st Qu.:1.000
##  Median :2.000   Median :1.00   Median :2.000   Median :2.000
##  Mean   :2.359   Mean   :1.05   Mean   :2.128   Mean   :1.675
##  3rd Qu.:3.000   3rd Qu.:1.33   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :3.00   Max.   :3.000   Max.   :3.000
##
##          gestat    brthwt       brthsco       medhist      stheight
##  38 Weeks+  :455   No :473   Min.   :0.0000   No :399   Min.   :-2.6400
##  34-37 Weeks: 43   Yes: 28   1st Qu.:0.0000   Yes:102   1st Qu.:-0.3900
```

```
##  <34 Weeks  :  3             Median :0.0000             Median : 0.1900
##                              Mean   :0.5289             Mean   : 0.2103
##                              3rd Qu.:1.0000             3rd Qu.: 0.7500
##                              Max.   :3.0000             Max.   : 3.2900
##
##     offschl          childint        parchcom        parpart
##  Min.   : 0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 3.000   1st Qu.:2.500   1st Qu.:5.000   1st Qu.:4.500
##  Median : 6.000   Median :3.250   Median :6.000   Median :5.500
##  Mean   : 8.884   Mean   :3.207   Mean   :6.106   Mean   :5.246
##  3rd Qu.:12.000   3rd Qu.:4.000   3rd Qu.:7.000   3rd Qu.:6.500
##  Max.   :65.000   Max.   :6.500   Max.   :8.000   Max.   :6.500
##
##     parschl           pvoc            pmat
##  Min.   :0.000   Min.   : 4.00   Min.   :13.0
##  1st Qu.:2.000   1st Qu.:43.00   1st Qu.:38.0
##  Median :3.000   Median :52.00   Median :45.0
##  Mean   :3.214   Mean   :52.48   Mean   :43.3
##  3rd Qu.:4.000   3rd Qu.:61.00   3rd Qu.:50.0
##  Max.   :6.000   Max.   :84.00   Max.   :60.0
##
```

It is useful to check the recoding and look for categories that have very low frequencies - e.g. moveschl and unempl might not be very helpful as 1 category very small. In general, there is likely more to do here, but this is quite a clean data set, so can move on.

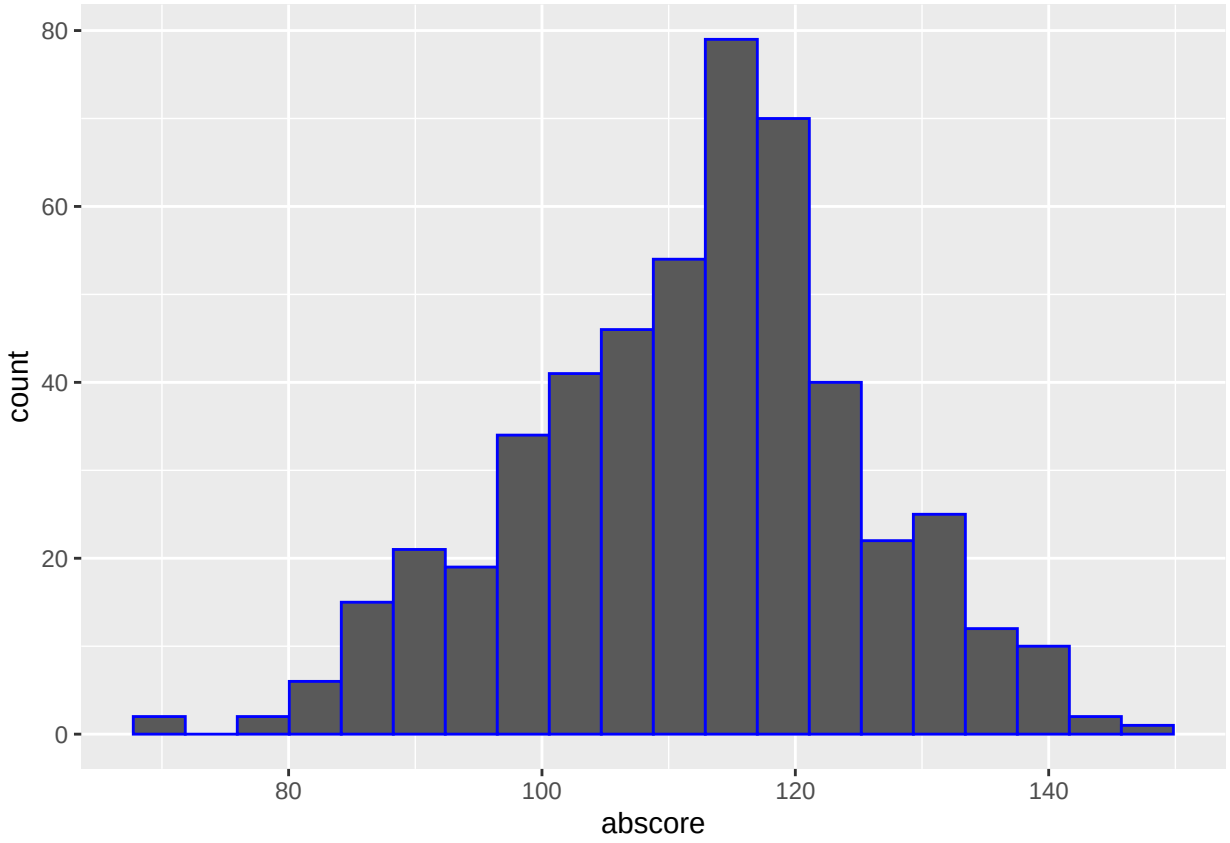
Exploratory Data Analysis

We're interested in possible relationships between sex, classyr, msoc (categorical) and bloodpb, ageint, famhist (continuous) on abscore, the educational attainment score for each child. We will use quite basic plots here for now, you might want to improve some of these for reporting purposes, but that's not necessary for basic checks at this stage.

Exploration of the abscore variable

First we can look at the overall distribution of the abscore variable.

```
# Overall distribution of abscore
Leadw %>%
  ggplot(aes(x=abscore)) +
  geom_histogram(bins=20,colour="blue")
```

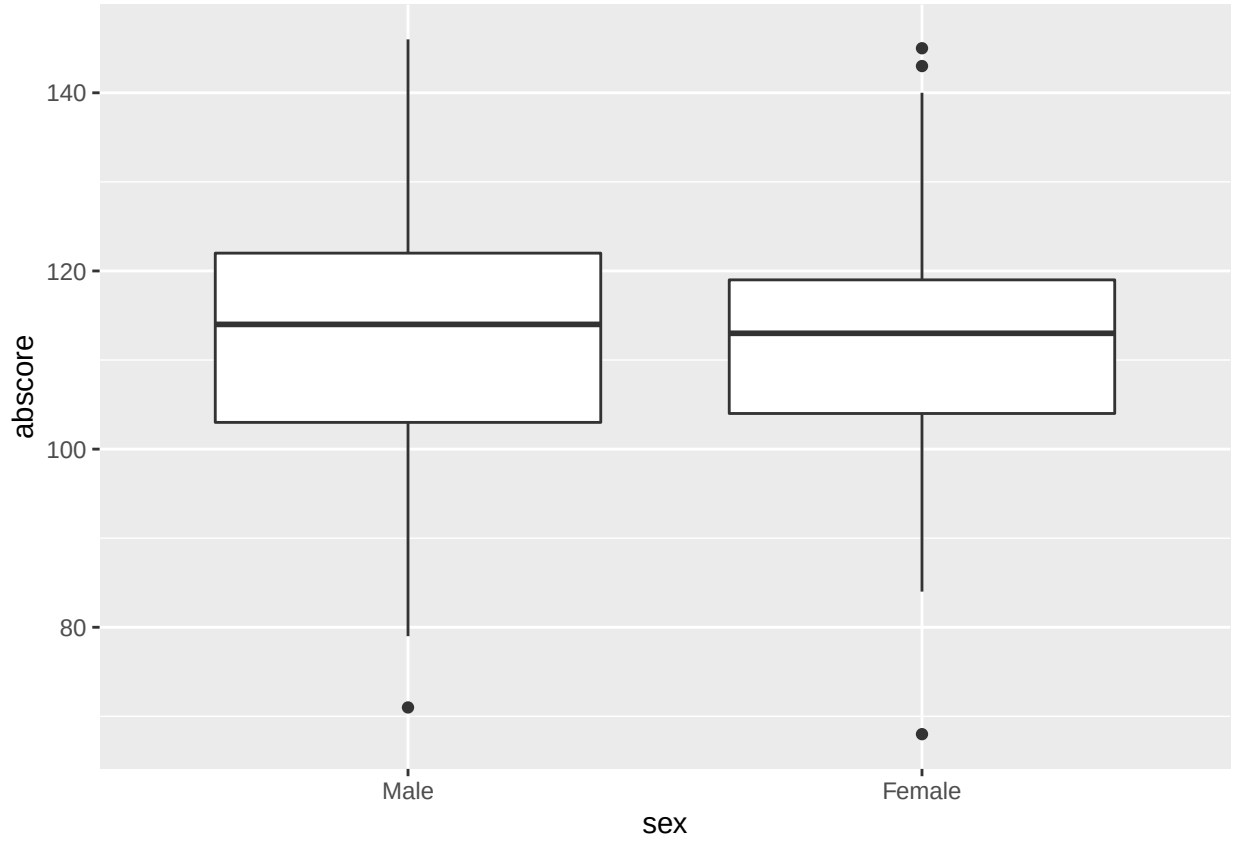


Look at abscore by each of the categorical variables

Now we can look at the distribution of abscore separated by each of the categorical variables, using grapical and numerical methods.

sex variable

```
Leadw %>%
  ggplot(aes(y=abscore,x=sex)) +
  geom_boxplot()
```

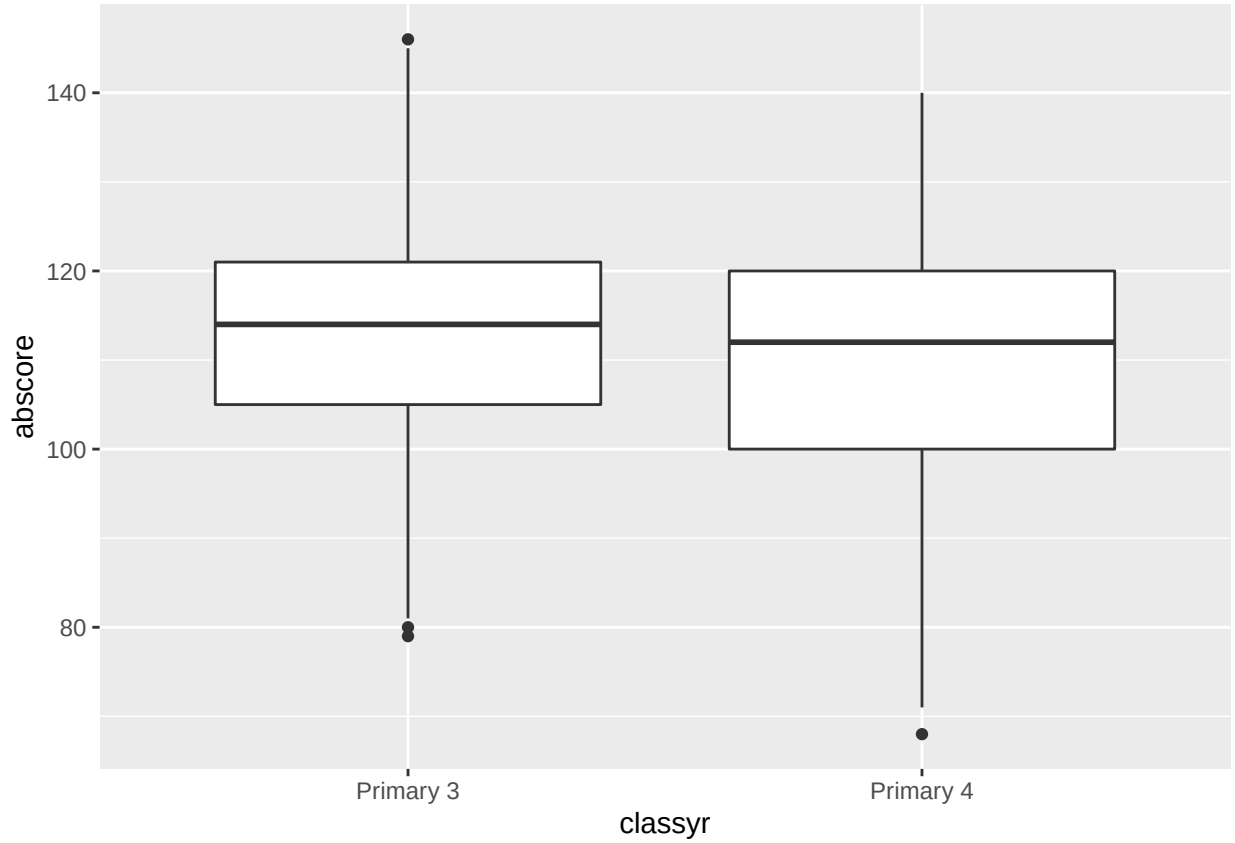



```
Leadw %>%
  group_by(sex) %>%
  summarize(min=min(abscore),
            mean=mean(abscore),
            median=median(abscore),
            max=max(abscore),
            SD=sd(abscore))
```

```
## # A tibble: 2 x 6
##   sex      min  mean median   max    SD
##   <fct>  <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Male      71  112.    114   146  14.1
## 2 Female    68  112.    113   145  12.5
```

classyr variable

```
Leadw %>%
  ggplot(aes(y=abscore,x=classyr)) +
  geom_boxplot()
```

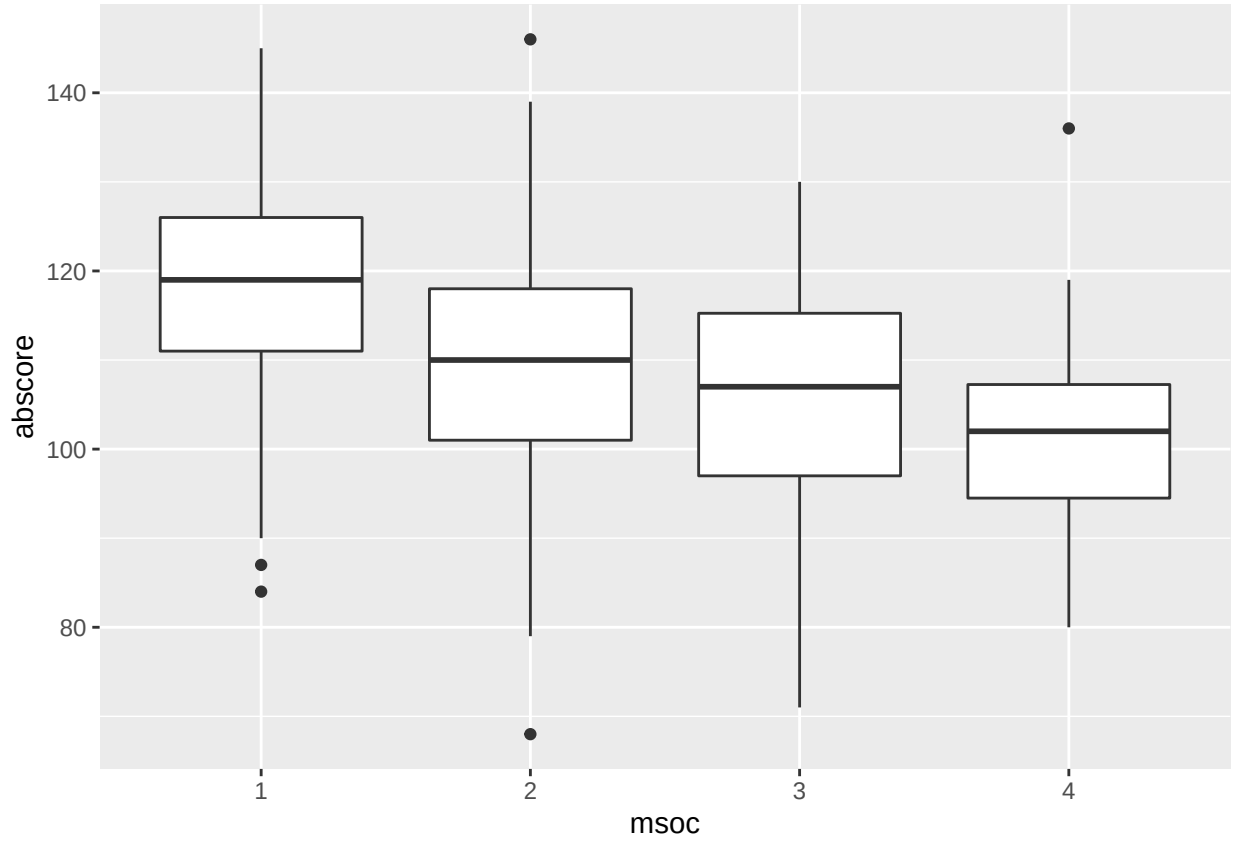



```
Leadw %>%
  group_by(classyr) %>%
  summarize(min=min(abscore),
            mean=mean(abscore),
            median=median(abscore),
            max=max(abscore),
            SD=sd(abscore))
```

```
## # A tibble: 2 x 6
##   classyr     min  mean median   max    SD
##   <fct>     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 Primary 3    79  113.    114   146  12.6
## 2 Primary 4    68  111.    112   140  14.0
```

msoc variable

```
Leadw %>%
  ggplot(aes(y=abscore,x=msoc)) +
  geom_boxplot()
```



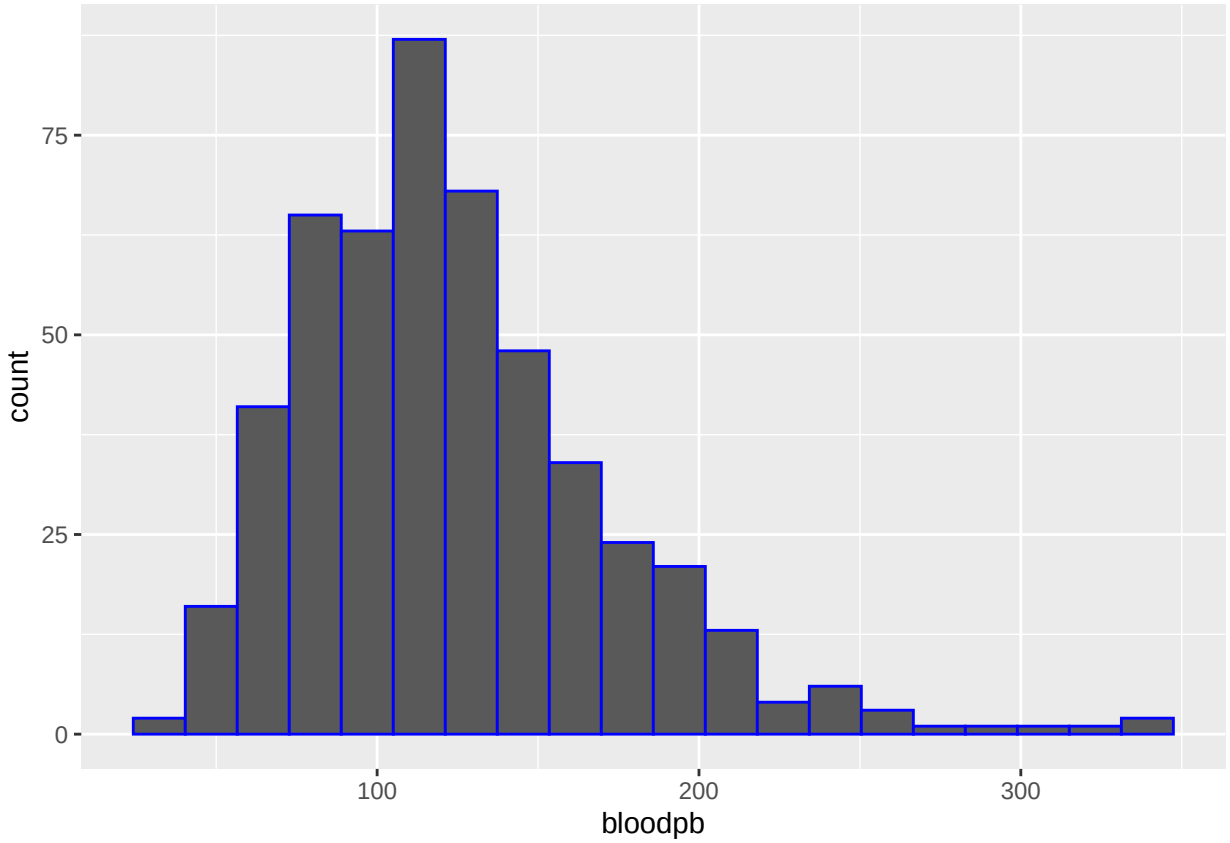

```
Leadw %>%
  group_by(msoc) %>%
  summarize(min=min(abscore),
            mean=mean(abscore),
            median=median(abscore),
            max=max(abscore),
            SD=sd(abscore))
```

```
## # A tibble: 4 x 6
##   msoc    min  mean median   max    SD
##   <fct> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 1        84  118.    119   145  11.5
## 2 2        68  109.    110   146  12.7
## 3 3        71  107.    107   130  13.9
## 4 4        80  102.    102   136  11.2
```

Look at individual continuous variables of interest

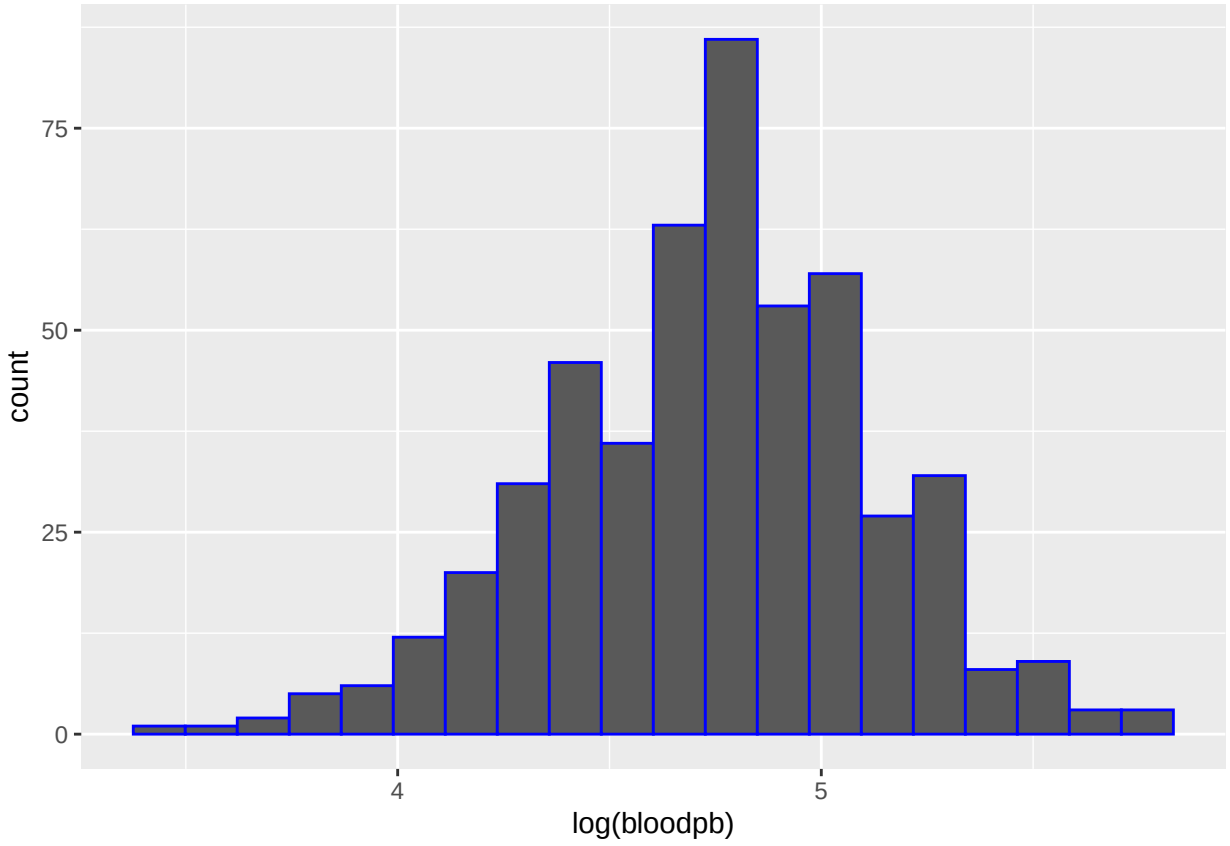
Then we can look at the overall distribution of each continuous variable of interest.

```
Leadw %>%
  ggplot(aes(x=bloodpb)) +
  geom_histogram(bins=20,colour="blue")
```

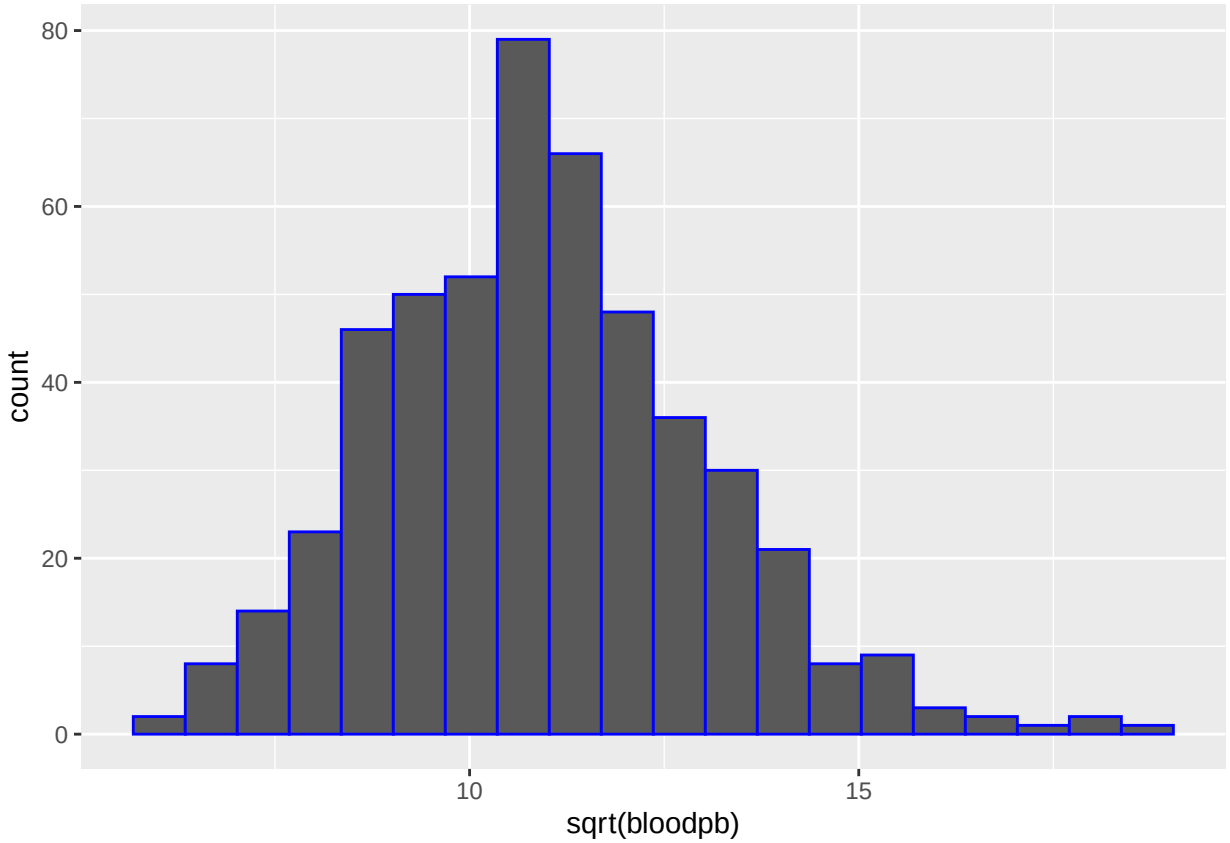


There is some positive skewness here (not unusually for this type of lab measurement). It may be easier to look at this on a transformed scale?

```
Leadw %>%
  ggplot(aes(x=log(bloodpb))) +
  geom_histogram(bins=20,colour="blue")
```

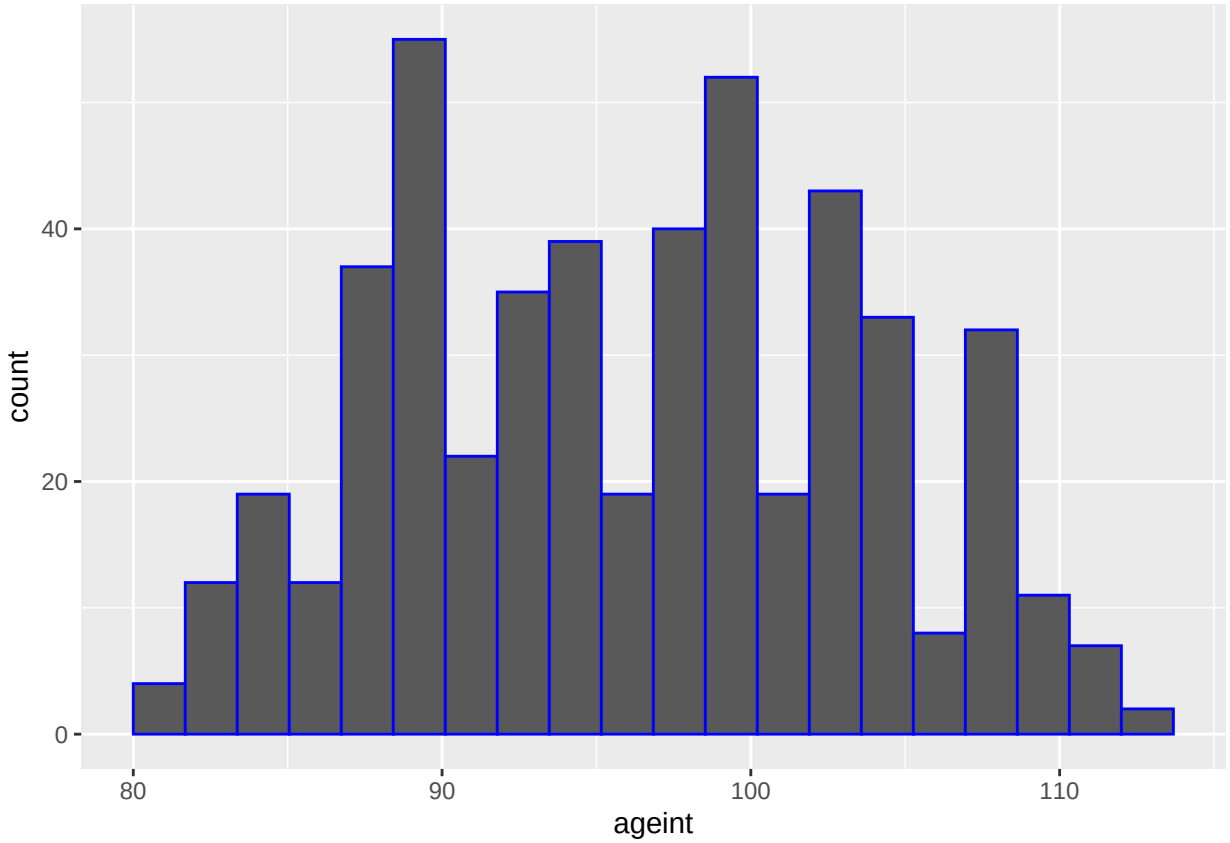



```
Leadw %>%
  ggplot(aes(x=sqrt(bloodpb))) +
  geom_histogram(bins=20,colour="blue")
```

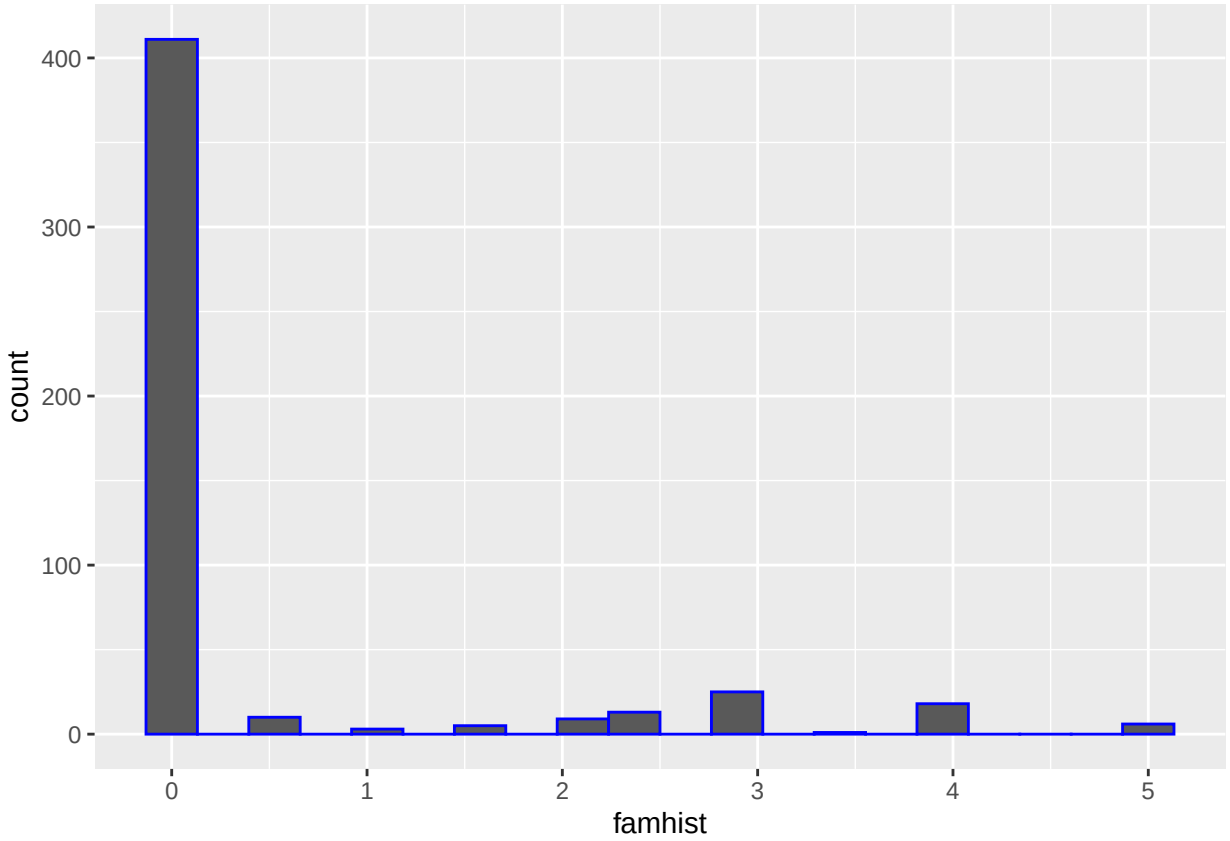


The square root scale may be a reasonable choice to make relationships easier to see.

```
Leadw %>%
  ggplot(aes(x=ageint)) +
  geom_histogram(bins=20,colour="blue")
```




```
Leadw %>%
  ggplot(aes(x=famhist)) +
  geom_histogram(bins=20,colour="blue")
```



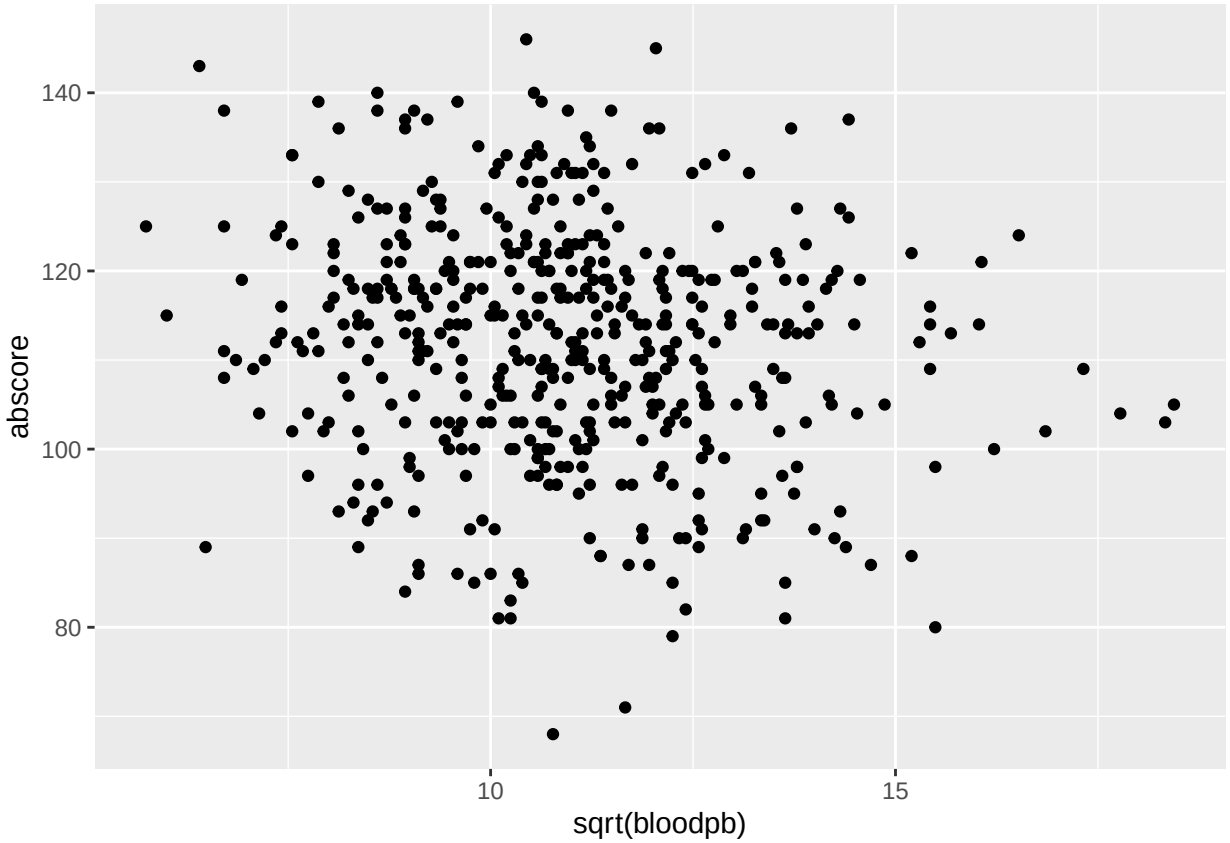
Here we note the irregular distribution, there are lots of cases with no family disruption, and zero or a small number of cases for some categories of family disruption. It would likely better be replaced by a classification of "None versus Some". We can move this variable into the categorical group for analysis.

```
Leadw<-Leadw %>%
  mutate(famhistbin=famhist!=0)     # Create Boolean variable - T = score of 1 or more, F = score of 0
```

We've now created a new Boolean variable named famihistbin, which is TRUE if famhist has a score of 1 or more, and FALSE if famhist has a score of 0.

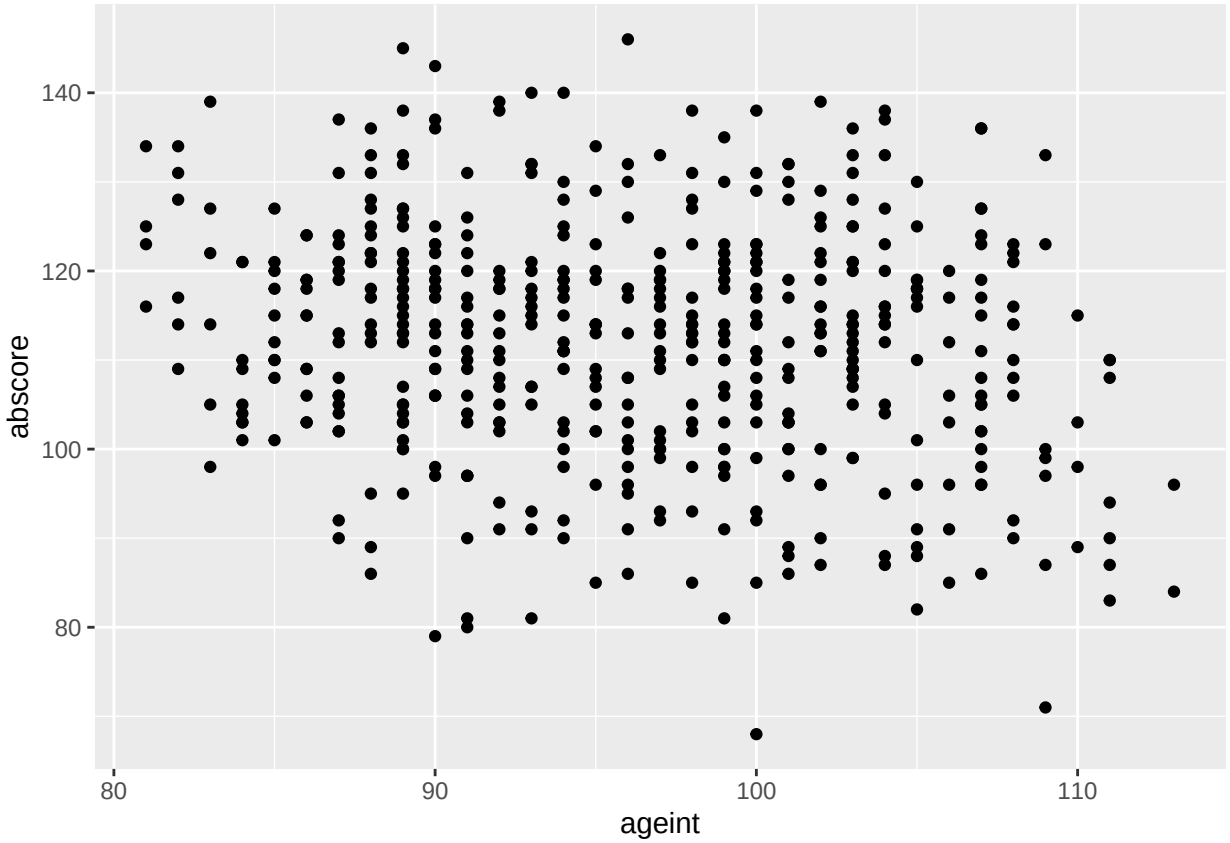
Look at scatterplots of continuous variables against abscore

```
Leadw %>%
  ggplot(aes(y=abscore,x=sqrt(bloodpb))) +
  geom_point()
```



A lower abscore seems to be associated with higher blood lead.

```
Leadw %>%
  ggplot(aes(y=abscore,x=ageint)) +
  geom_point()
```



This plot is slightly difficult to interpret, we will need some quantification to really see if there is any relationship.

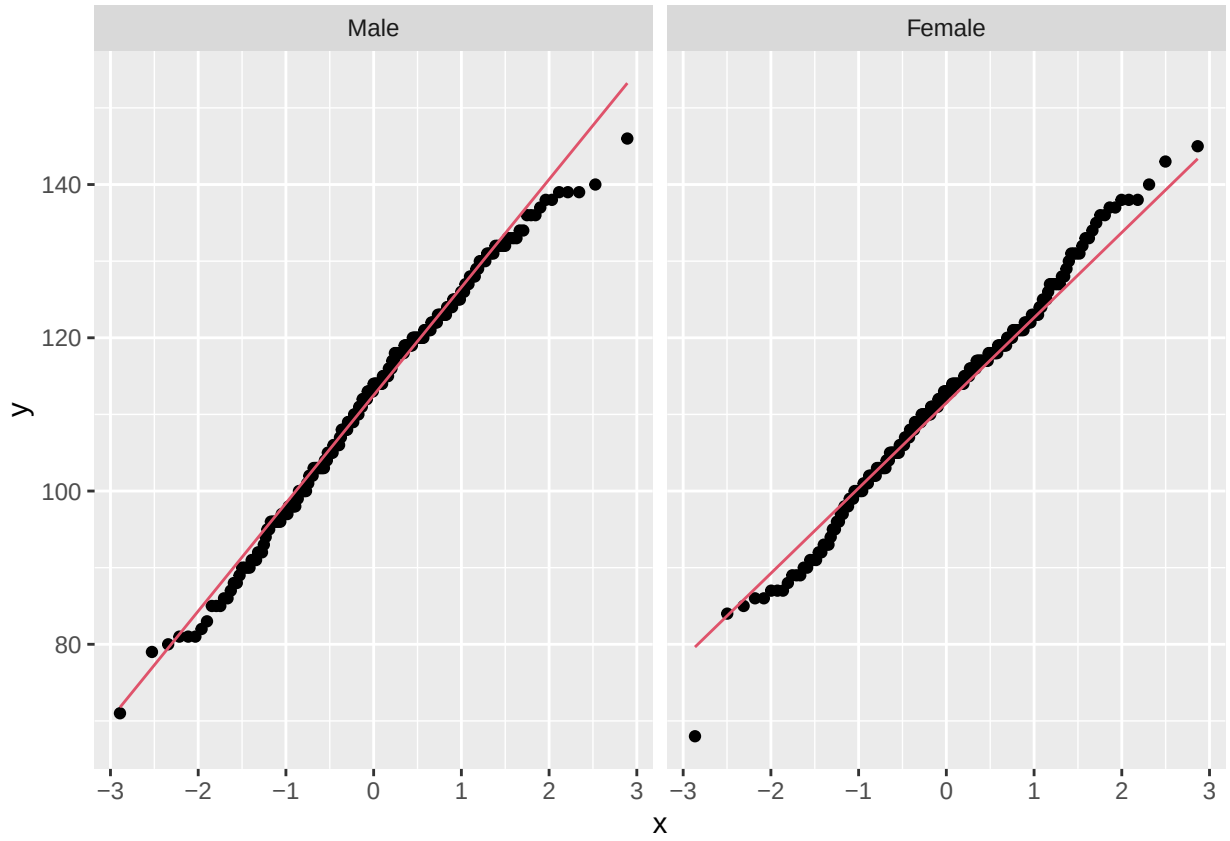
Develop the Analysis Plan

This requires thinking time, it happens off-line, but we will need to think about the data types, the structure of the problem, etc. Here we can use t-tests to assess abscore vs categorical variables. msoc would need to use a 1-way ANOVA (discussed later, so leave for now). The association between abscore and the continuous variables can be assessed by correlation coefficient, and later we would start using regression models, etc. to do more complex multifactorial investigation.

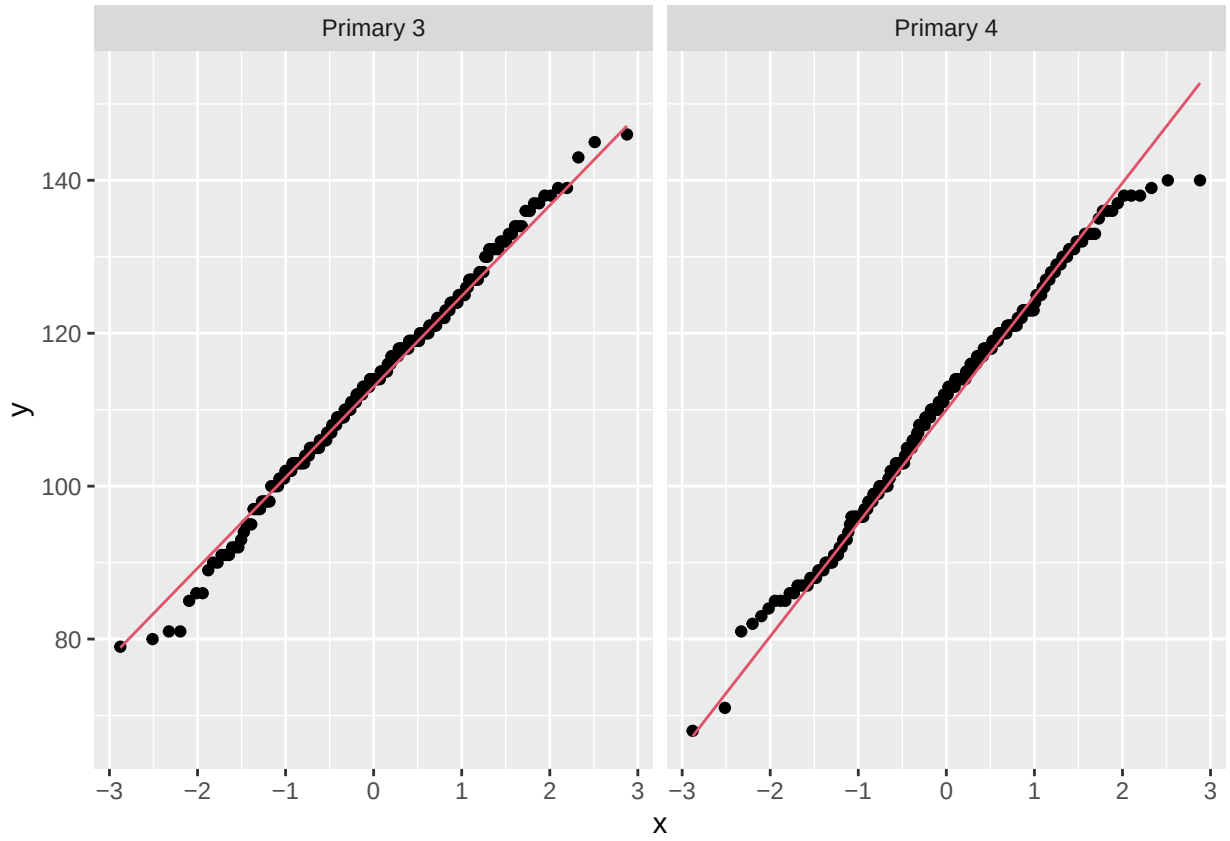
Investigate Assumptions

We can see SDs for abscore in different groups from above, so we just need to add Q_Q plots here.

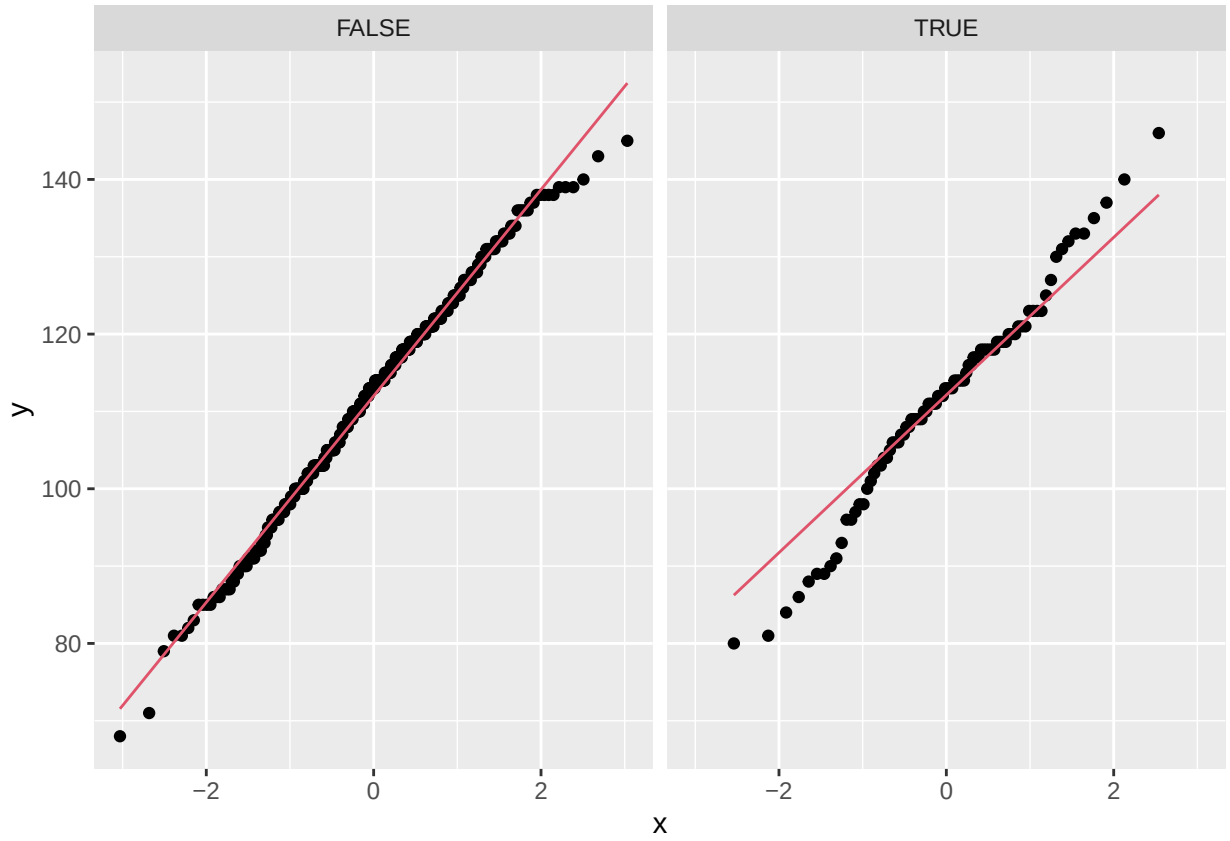
```
Leadw %>%
  ggplot(aes(sample=abscore)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_grid(cols = vars(sex))
```




```
Leadw %>%
  ggplot(aes(sample=abscore)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_grid(cols = vars(classyr))
```




```
Leadw %>%
  ggplot(aes(sample=abscore)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_grid(cols = vars(famhistbin))
```



The above sets of plots look reasonably Normal, so there are no particular concerns regarding the t-test assumptions. The poorest fir to the normal distribution is the smaller group of children with family history of disruption. This is expected as smaller samples have more noisy data, so we are less concerned about this.

Carry Out Analysis

t-tests

```
Leadw %>% t.test(.$abscore~.$sex,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  .$abscore by .$sex
## t = 0.097503, df = 498.39, p-value = 0.9224
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -2.22414  2.45642
## sample estimates:
##   mean in group Male mean in group Female
##             112.0536             111.9375
```

```
Leadw %>% t.test(.$abscore~.$classyr,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  .$abscore by .$classyr
## t = 2.1779, df = 494.35, p-value = 0.02989
## alternative hypothesis: true difference in means between group Primary 3 and group Primary 4 is not
## 95 percent confidence interval:
##  0.2534943 4.9282804
## sample estimates:
## mean in group Primary 3 mean in group Primary 4
##                113.3012                110.7103
```

```
Leadw %>% t.test(.$abscore~.$famhistbin,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  .$abscore by .$famhistbin
## t = 0.18365, df = 132.92, p-value = 0.8546
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to
## 95 percent confidence interval:
##  -2.755152  3.319142
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            112.0487            111.7667
```

We can omit the ANOVA for msoc for now.

Correlation tests

```
with(Leadw,cor(abscore,sqrt(bloodpb)))  # Simple base R function - easier to use directly here, rather
```

```
## [1] -0.1557688
```

```
with(Leadw,cor(abscore,ageint))
```

```
## [1] -0.1778181
```

Both bloodpb and age seem to be associated with decreasing abscore, which is interesting. More analysis will likely continue from here.

Interpretation

Here you will interpret the results of each method and plot and explain how they answer the questions.

Conclusions

This is an overall conclusion of all the findings, probably highlighting the bigger effects found.