Code

Assignment 1

Gus Lipkin ~ glipkin6737@floridapoly.edu (mailto:glipkin6737@floridapoly.edu)

Problem 8

This exercise relates to the `College` data set, which can be found in the file `College.csv` on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are

[ommitted for simplicity]

Before reading the data into `R, it can be viewed in Excel or a text editor.

8a

Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

Hide

```
college <- read.csv("College.csv")
```

8b

Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
rownames(college) <- college[, 1]

View(college)
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
college <- college[, -1]

View(college)
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

Hide

```
# Commenting this out because it doesn't work with html notebooks
# View(college)
head(college)
```

X <chr>	Private <chr>	A... <int>	Acc... <int>	Enroll <int>	Top10perc <int>	Top25perc <int>	F.Und
1 Abilene Christian University	Yes	1660	1232	721	23	52	
2 Adelphi University	Yes	2186	1924	512	16	29	
3 Adrian College	Yes	1428	1097	336	22	50	
4 Agnes Scott College	Yes	417	349	137	60	89	
5 Alaska Pacific University	Yes	193	146	55	16	44	

6	Albertson College	Yes	587	479	158	38	62

6 rows | 1-9 of 19 columns

Hide

```
# Using row names is outdated...
rownames(college) <- college[, 1]
# View(college)
head(college)
```

	X <chr>	Private <chr>	A... <int>	Acc... <int>	Er <
Abilene Christian University	Abilene Christian University	Yes	1660	1232	
Adelphi University	Adelphi University	Yes	2186	1924	
Adrian College	Adrian College	Yes	1428	1097	
Agnes Scott College	Agnes Scott College	Yes	417	349	
Alaska Pacific University	Alaska Pacific University	Yes	193	146	
Albertson College	Albertson College	Yes	587	479	

6 rows | 1-7 of 19 columns

Hide

```
college <- college[, -1]
# View(college)
head(college)
```

	Private <chr>	A... <int>	Acc... <int>	Enroll <int>	Top10perc <int>	Top25perc <int>	F.Unde
Abilene Christian University	Yes	1660	1232	721	23	52	
Adelphi University	Yes	2186	1924	512	16	29	
Adrian College	Yes	1428	1097	336	22	50	
Agnes Scott College	Yes	417	349	137	60	89	
Alaska Pacific University	Yes	193	146	55	16	44	
Albertson College	Yes	587	479	158	38	62	

6 rows | 1-8 of 18 columns

8c

8c i

> Use the `summary()` function to produce a numerical summary of the variables in the data set.

Hide

```
summary(college)
```

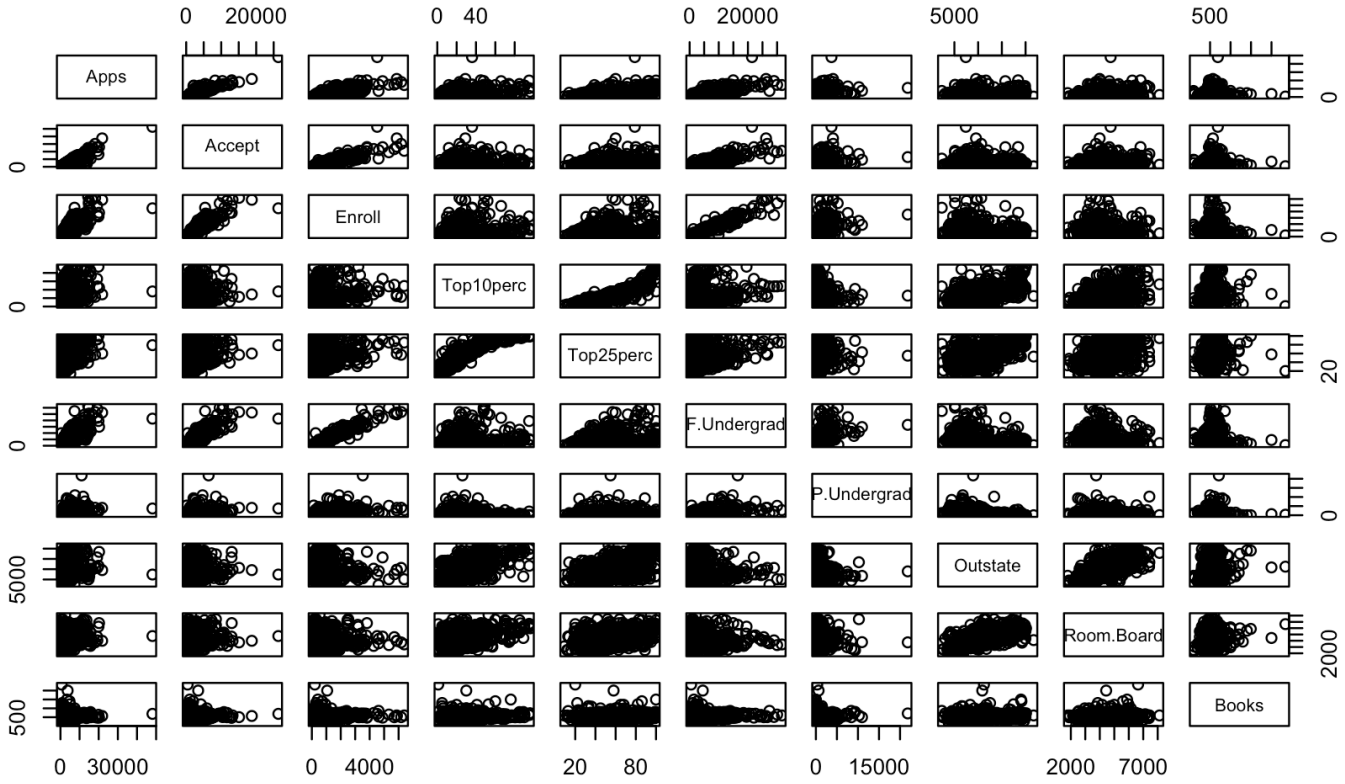
```
   Private                Apps           Accept          Enroll       Top10perc
Top25perc      F.Undergrad
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Mi
n.   :  9.0   Min.   :  139
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1s
t Qu.: 41.0   1st Qu.:  992
 Mode  :character   Median : 1558   Median : 1110   Median : 434   Median :23.00   Me
dian : 54.0   Median : 1707
                    Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Me
an   : 55.8   Mean   : 3700
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3r
d Qu.: 69.0   3rd Qu.: 4005
                    Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Ma
x.   :100.0   Max.   :31643
  P.Undergrad         Outstate       Room.Board       Books           Personal
PhD            Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min
.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st
Qu.: 62.00   1st Qu.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Med
ian : 75.00   Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mea
n   : 72.66   Mean   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd
Qu.: 85.00   3rd Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max
.   :103.00   Max.   :100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

8c ii

Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

Hide

```
# The first column is whether or not it is a private school (character) and isn't all
owed by `pairs`
pairs(college[,2:11])
```

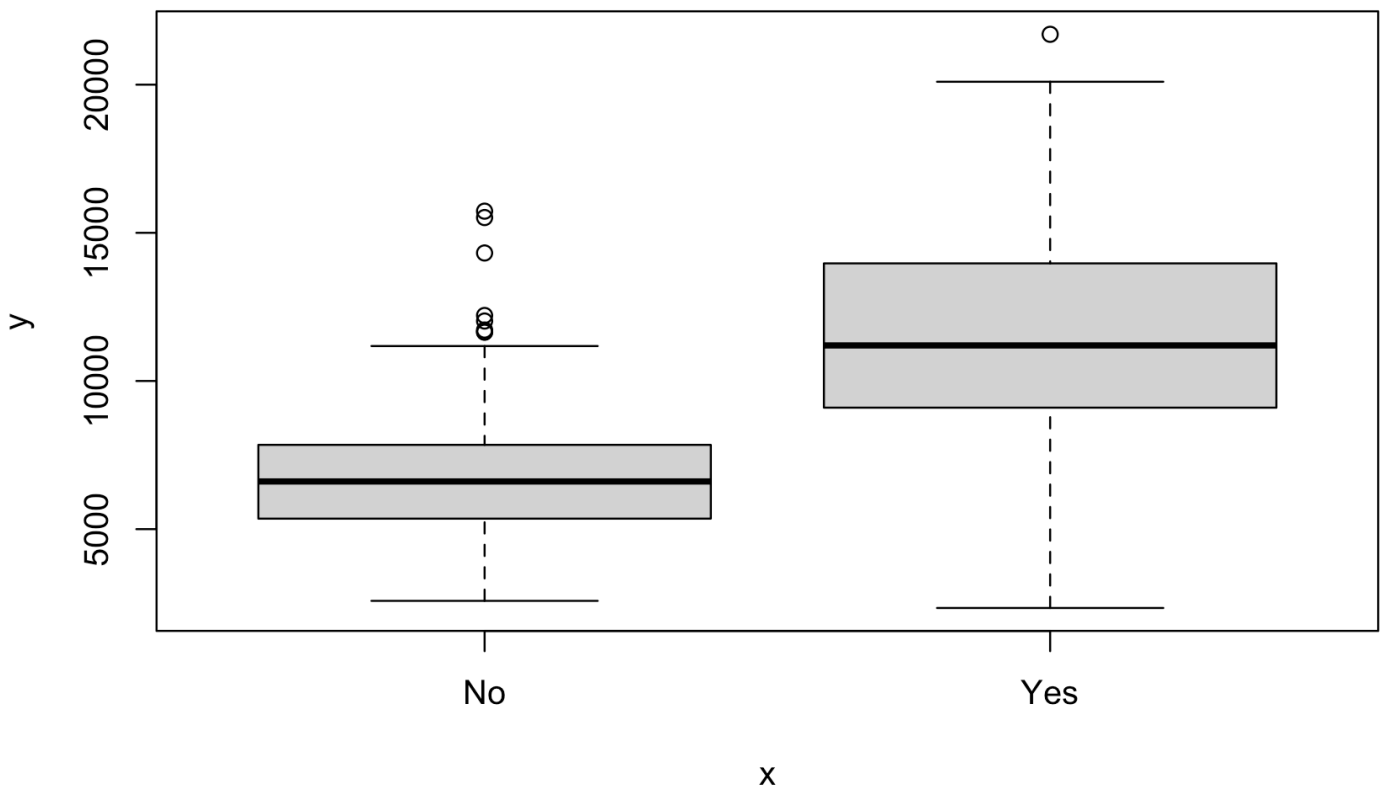


8c iii

Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

Hide

```
college$Private <- as.factor(college$Private)
plot(college$Private, college$Outstate)
```



8c iv

Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. `Elite <- rep("No", nrow(college))`

```
Elite[college$Top10perc > 50] <- "Yes"
```

```
Elite <- as.factor(Elite)
```

`college <- data.frame(college, Elite)` Use the `summary()` function to see how many elite univer- sities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

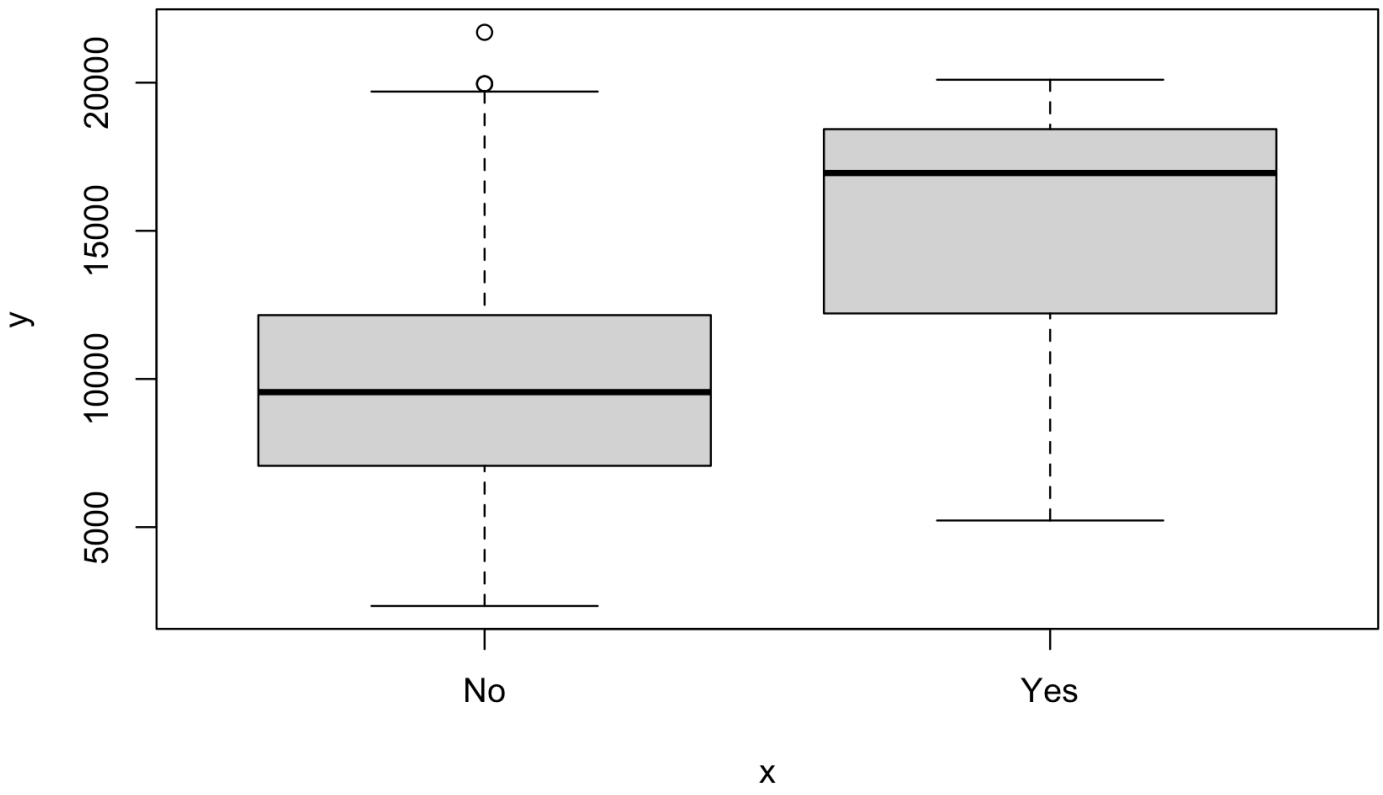
Hide

```
# Here's the code they gave you, but it's super inefficient
# Elite <- rep("No", nrow(college))
# Elite[college$Top10perc > 50] <- "Yes"
# Elite <- as.factor(Elite)
# college <- data.frame(college, Elite)
college$Elite <- as.factor(ifelse(college$Top10perc > 50, "Yes", "No"))
summary(college$Elite)
```

```
 No Yes
699  78
```

Hide

```
plot(college$Elite, college$Outstate)
```



8c v

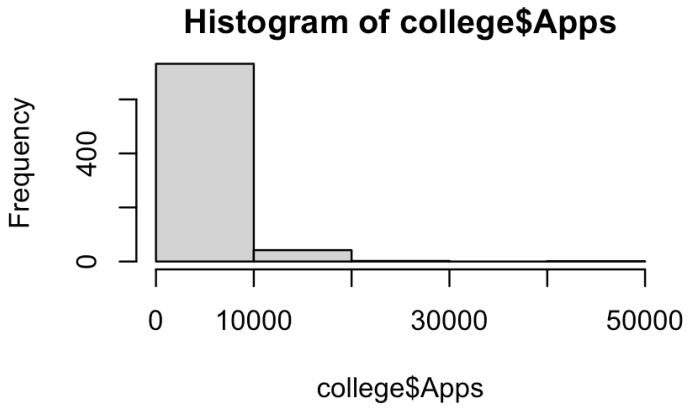
Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative vari- ables. You may find the command `par(mfrow = c(2, 2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

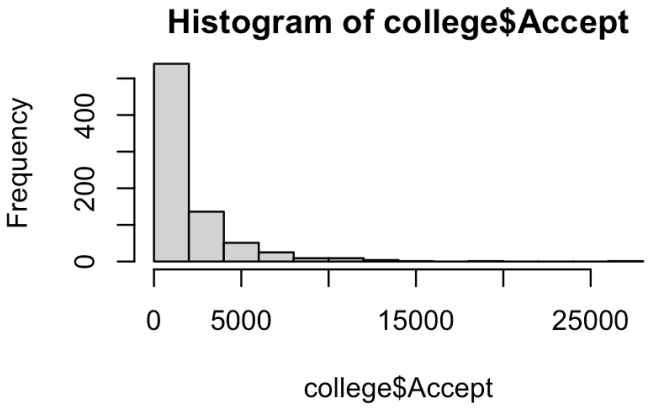
Hide

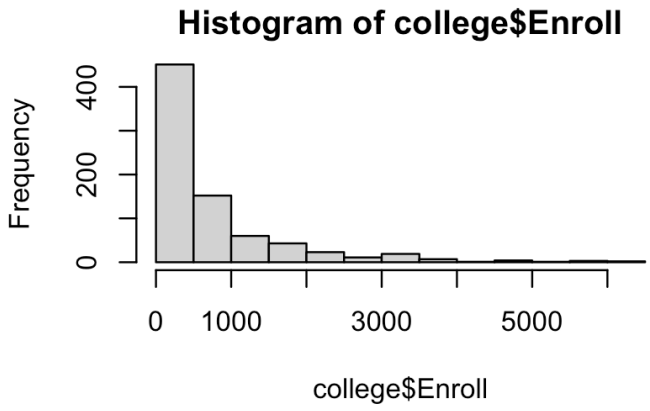
```
par(mfrow = c(2, 2))
hist(college$Apps, breaks = 5)
hist(college$Accept, breaks = 10)
```

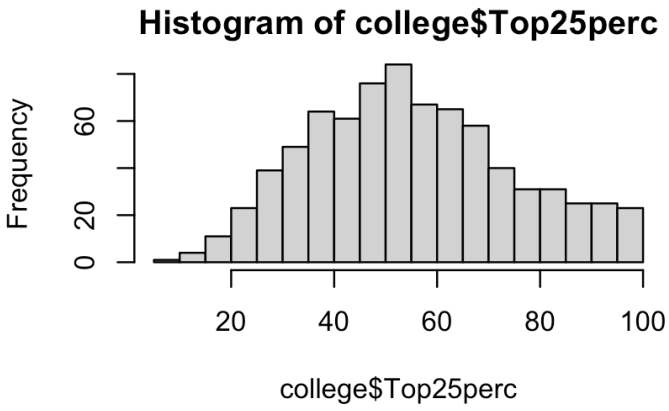
Hide

```
hist(college$Enroll, breaks = 15)
hist(college$Top25perc, breaks = 25)
```









8c vi

Continue exploring the data, and provide a brief summary of what you discover.

Hide

```
college$AcceptRate <- college$Accept / college$Apps
acceptRate <- lm(Grad.Rate ~ AcceptRate, data = college)
acceptRate
```

```

Call:
lm(formula = Grad.Rate ~ AcceptRate, data = college)

Coefficients:
(Intercept)   AcceptRate
      90.49       -33.51
```

Hide

```
summary(acceptRate)
```

```

Call:
lm(formula = Grad.Rate ~ AcceptRate, data = college)

Residuals:
    Min      1Q  Median      3Q     Max
-58.491 -10.806   0.968  12.496  57.411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   90.493      3.059   29.58  < 2e-16 ***
AcceptRate   -33.510      4.018   -8.34 3.39e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.47 on 775 degrees of freedom
Multiple R-squared:  0.08235,   Adjusted R-squared:  0.08117
F-statistic: 69.55 on 1 and 775 DF,  p-value: 3.39e-16
```

People generally assume that schools with lower acceptance rates are more prestigious and more academically rigorous. I wanted to see if that is true. I created a metric for the acceptance rate: $\frac{\text{\# accepted}}{\text{\# apps}}$. I then performed a simple linear regression to see if the acceptance rate is a good predictor of the graduation rate. In theory, it is based on p-value, but not based on the R-squared.

9

> This exercise involves the `Auto` data set studied in the lab. Make sure that the missing values have been removed from the data.

Hide

```
library(data.table)
library(tidyverse)
auto <- fread("Auto.csv")
# This checks for na values
sum(is.na(auto))
```

```
[1] 0
```

9a

> Which of the predictors are quantitative, and which are qualitative?

Hide

```
# These are quantitative
colnames(select_if(auto, is.numeric))
```

```
[1] "mpg"          "cylinders"    "displacement" "weight"       "acceleration" "year"
"origin"
```

Hide

```
# These are qualitative
colnames(select_if(auto, negate(is.numeric)))
```

```
[1] "horsepower" "name"
```

9b

What is the range of each quantitative predictor? You can answer this using the `range()` function.

Hide

```
sapply(select_if(auto, is.numeric), range)
```

```
     mpg cylinders displacement weight acceleration year origin
[1,]  9.0         3           68   1613          8.0   70      1
[2,] 46.6         8          455   5140         24.8   82      3
```

9c

What is the mean and standard deviation of each quantitative predictor?

Hide

```
sapply(select_if(auto, is.numeric), mean)
```

```
        mpg    cylinders displacement       weight acceleration         year       o
rigin
   23.515869     5.458438   193.532746  2970.261965    15.555668    75.994962      1.5
74307
```

Hide

```
sapply(select_if(auto, is.numeric), sd)
```

```
        mpg    cylinders displacement       weight acceleration         year       o
rigin
   7.8258039    1.7015770  104.3795833  847.9041195    2.7499953    3.6900049    0.80
25495
```

9d

Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

Hide

```
smallAuto <- auto[c(1:9, 85:nrow(auto)),]
summary(select_if(smallAuto, is.numeric))
```

```
      mpg          cylinders      displacement       weight      acceleration
year           origin
 Min.   :11.00   Min.   :3.000   Min.   : 68.0   Min.   :1649   Min.   : 8.50   Min.
:70.00   Min.   :1.000
 1st Qu.:18.00   1st Qu.:4.000   1st Qu.: 98.5   1st Qu.:2211   1st Qu.:14.00   1st Q
u.:75.00   1st Qu.:1.000
 Median :23.95   Median :4.000   Median :144.5   Median :2792   Median :15.55   Media
n :77.00   Median :1.000
 Mean   :24.45   Mean   :5.366   Mean   :186.8   Mean   :2931   Mean   :15.73   Mean
:77.14   Mean   :1.602
 3rd Qu.:30.65   3rd Qu.:6.000   3rd Qu.:250.0   3rd Qu.:3494   3rd Qu.:17.27   3rd Q
u.:80.00   3rd Qu.:2.000
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :4997   Max.   :24.80   Max.
:82.00   Max.   :3.000
```

9e

> Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

Hide

```
SmartEDA::ExpReport(data = auto, op_file = "9e.html", Target = "mpg")
```

Exploratory Data Analysis Report for the Boston Dataset with a Focus on Gas Mileage

Exploratory Data Analysis Report

Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

Overview of the data

Descriptions <chr>	**Value** <chr>
Sample size (nrow)	397
No. of variables (ncol)	9
No. of numeric/interger variables	7
No. of factor variables	0
No. of text variables	2
No. of logical variables	0

No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	100% (9)

1-10 of 13 rows Previous **1** 2 Next

Structure of the data

Ind...	Variable_Name	Variable_Type	Sampl...	Missing_Count	Per_of_Missing	No_of_
<dbl>	<chr>	<chr>	<int>	<int>	<dbl>	
1	mpg	numeric	397	0	0	
2	cylinders	integer	397	0	0	
3	displacement	numeric	397	0	0	
4	horsepower	character	397	0	0	
5	weight	integer	397	0	0	
6	acceleration	numeric	397	0	0	
7	year	integer	397	0	0	
8	origin	integer	397	0	0	
9	name	character	397	0	0	

9 rows

Target variable

Summary of continuous dependent variable

1. Variable name - **mpg**
2. Variable description - ****

2. Summary of numerical variables

Summary statistics when dependent variable is Continuous **mpg**.

```
ExpNumStat(data,by="A",gp=Target,Qnt=seq(0,1,0.1),MesofShape=2,Outlier=TRUE,round=2)
```

Vname	Group	Note	TN	n...	nZero	n...	NegInf	PosInf	NA_Value
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

acceleration	mpg	Cor b/w mpg	397	0	0	397	0	0	0
displacement	mpg	Cor b/w mpg	397	0	0	397	0	0	0
mpg	mpg	Cor b/w mpg	397	0	0	397	0	0	0
weight	mpg	Cor b/w mpg	397	0	0	397	0	0	0
year	mpg	Cor b/w mpg	397	0	0	397	0	0	0

5 rows | 1-10 of 36 columns

3. Distributions of numerical variables

Graphical representation of all numeric features, used below types of plots to explore the data

- Quantile-quantile plot (Univariate)
- Density plot (Univariate)
- Scatter plot (Bivariate)

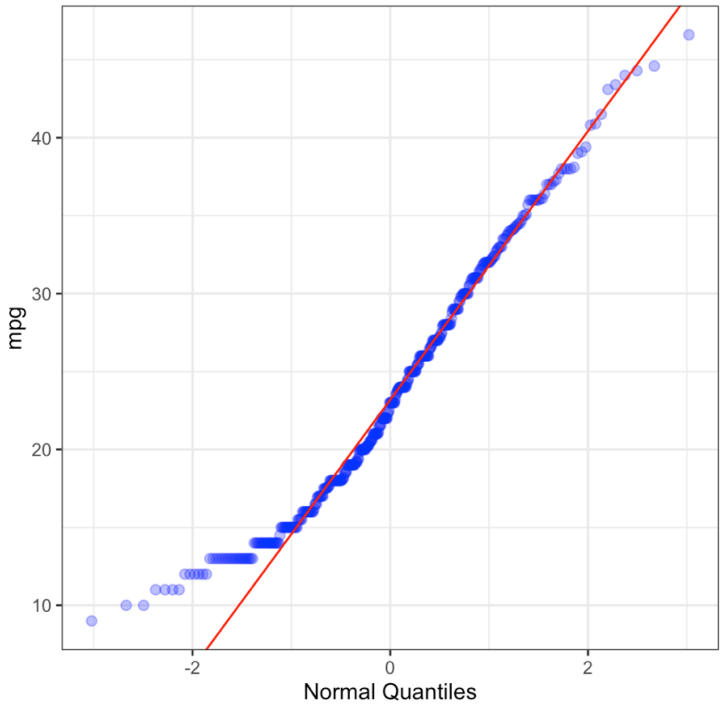
Quantile-quantile plot for Numerical variables - Univariate

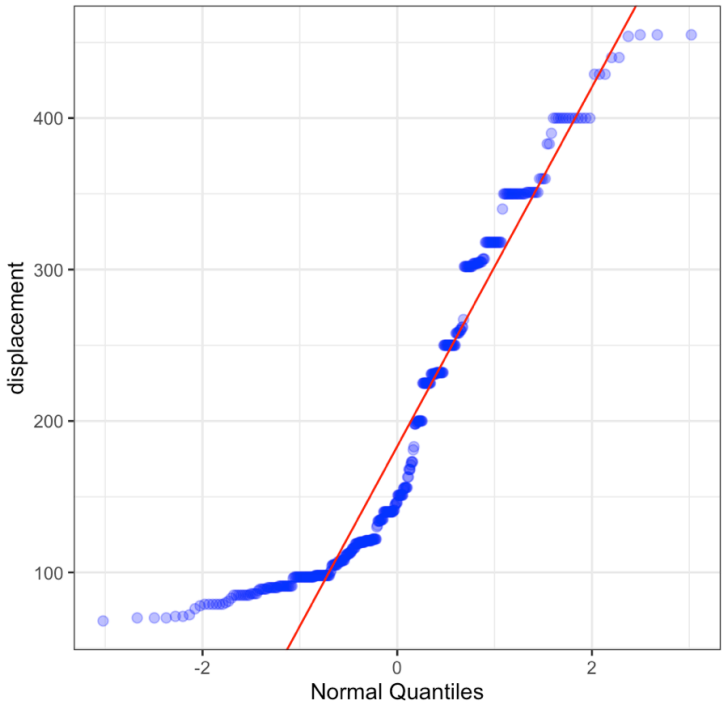
Quantile-quantile plot for all Numerical variables

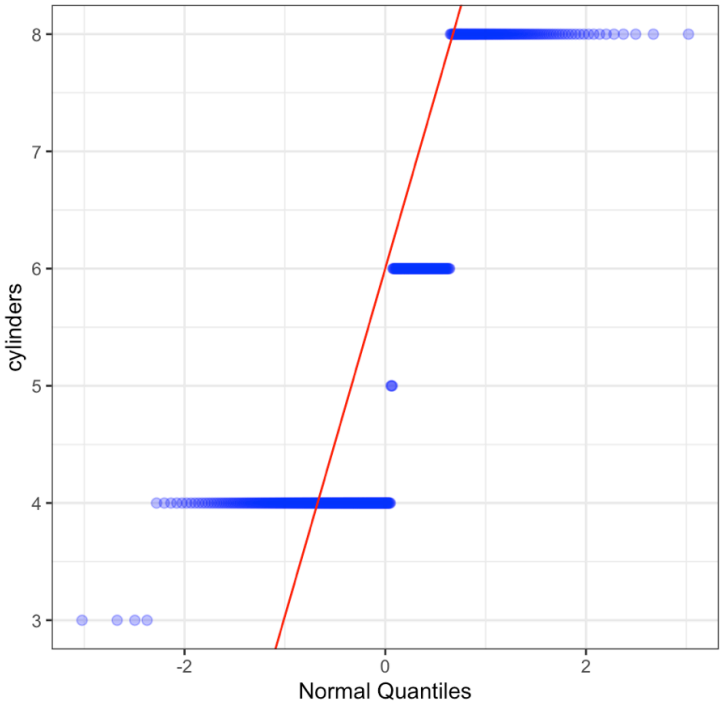
```
ExpOutQQ(data,nlim=4,fname=NULL,Page=c(2,2),sample=sn)
```

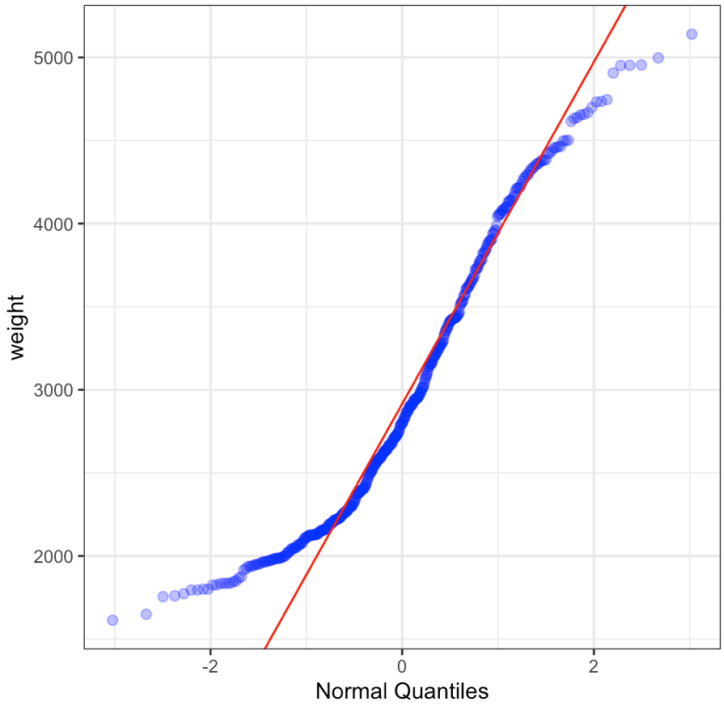
```
## $`0`
```

page 1 of 2

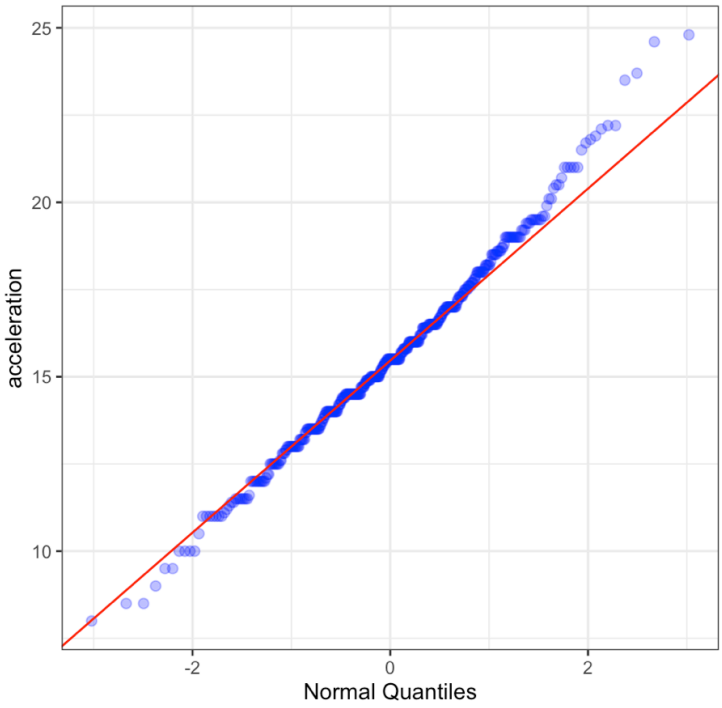

mpg
40
30
20
10
-2
0
2
Normal Quantiles

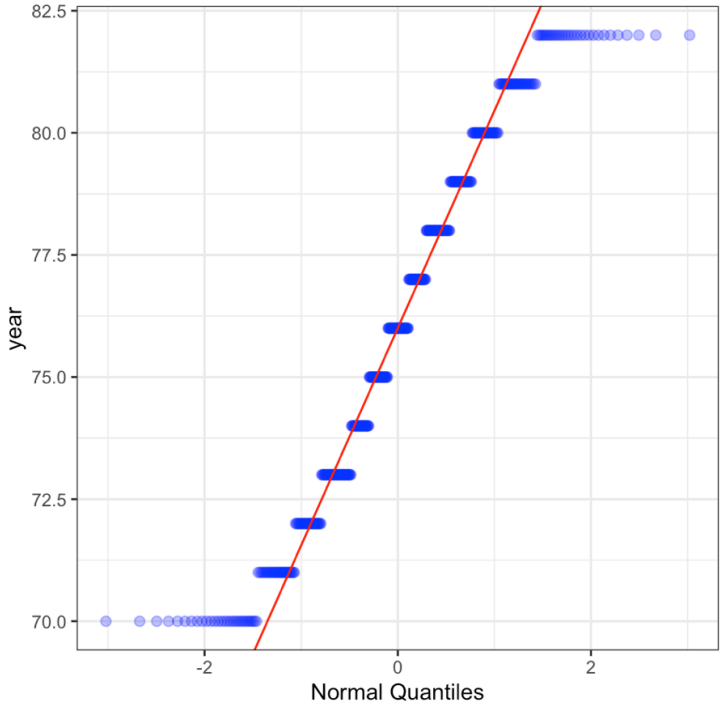

displacement
400
300
200
100
-2
0
2
Normal Quantiles


cylinders
8
7
6
5
4
3
-2
0
2
Normal Quantiles


weight
5000
4000
3000
2000
-2
0
2
Normal Quantiles

page 2 of 2





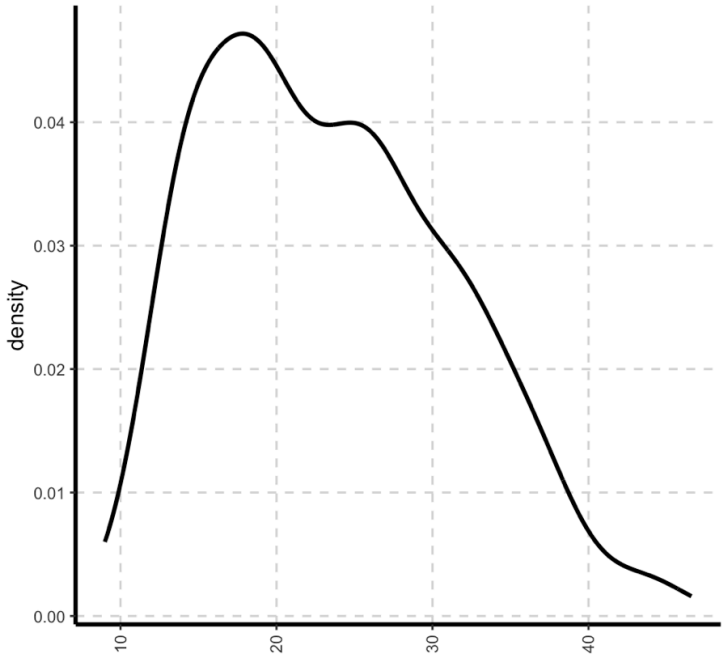
Density plots for numerical variables - Univariate

Density plot for all numerical variables

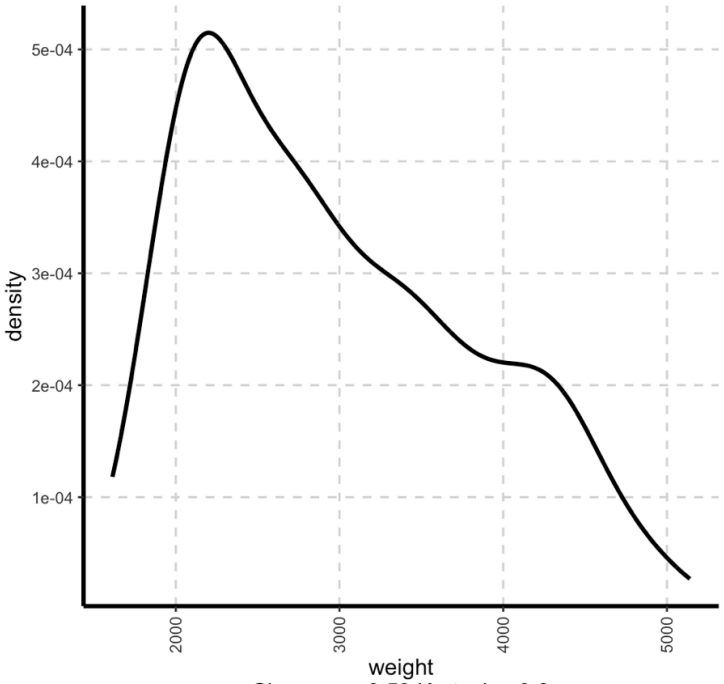
```
ExpNumViz(data,target=NULL,nlim=10,fname=NULL,col=NULL,theme=theme,Page=c(2,2),sample
=sn)
```

$`0`

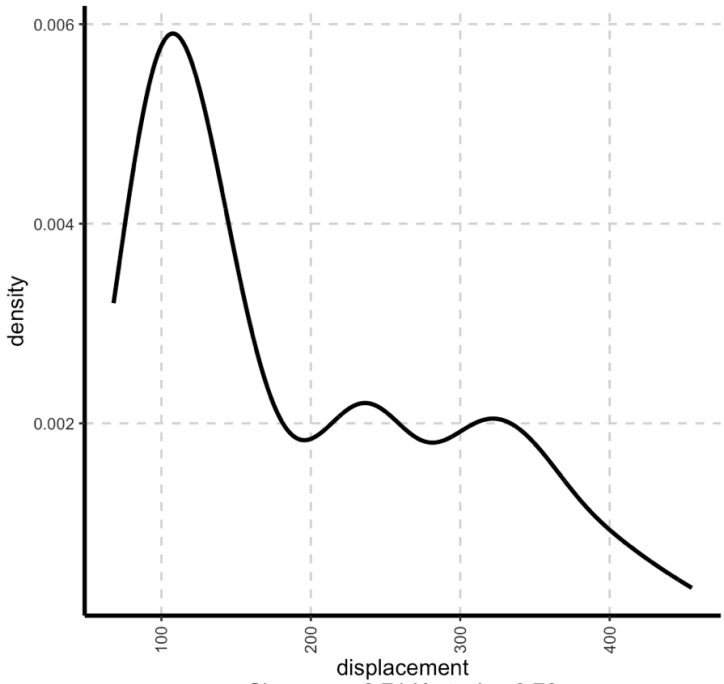
page 1 of 2



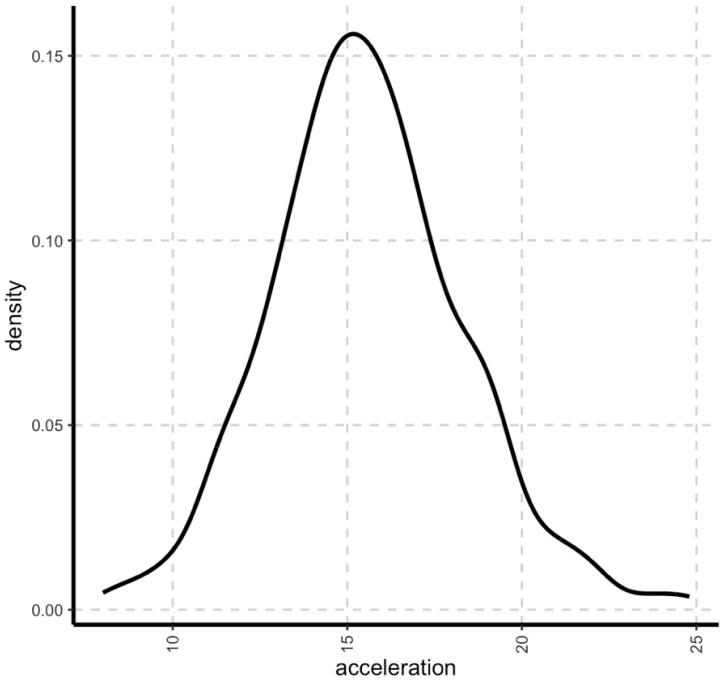
Skewness: 0.45 Kurtosis: -0.53



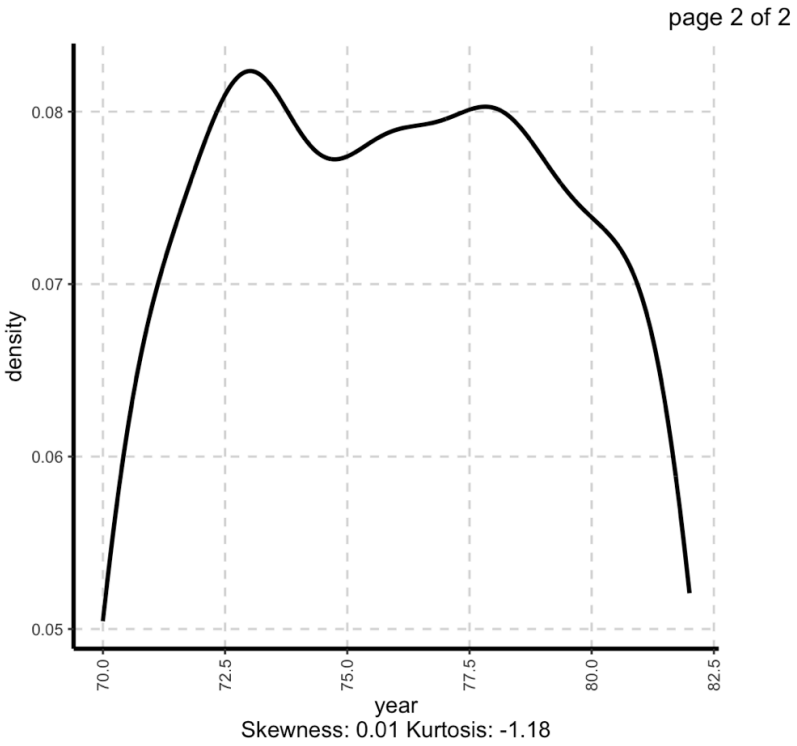
Skewness: 0.53 Kurtosis: -0.8



Skewness: 0.71 Kurtosis: -0.76



Skewness: 0.28 Kurtosis: 0.42

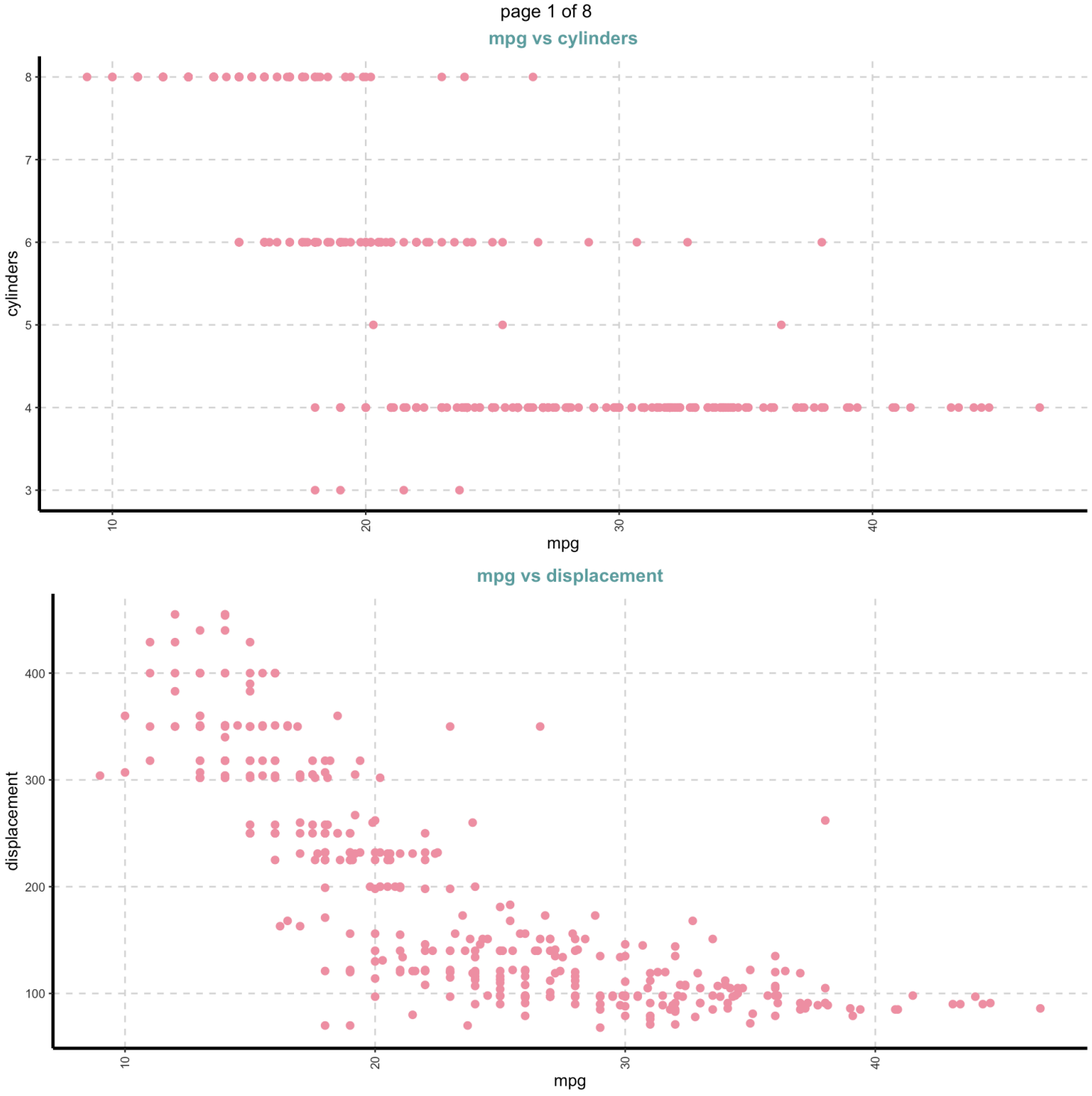


Scatter plot for all Numeric variables

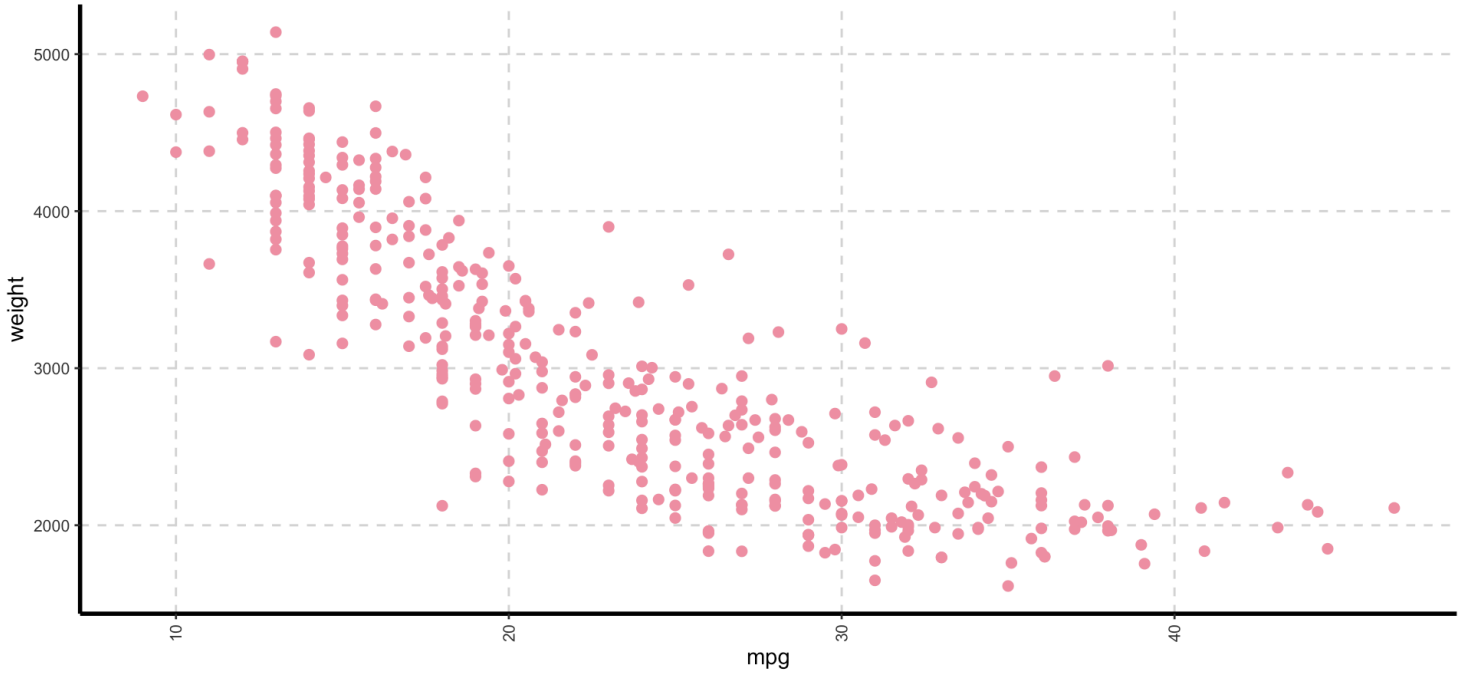
Scatter plot between all numeric variables and target variable **mpg**. This plot help to examine how well a target variable is correlated with list of dependent variables in the data set.

```
ExpNumViz(data,target=NULL,nlim=5,Page=c(2,1),theme=theme,sample=sn,scatter=TRUE)
```

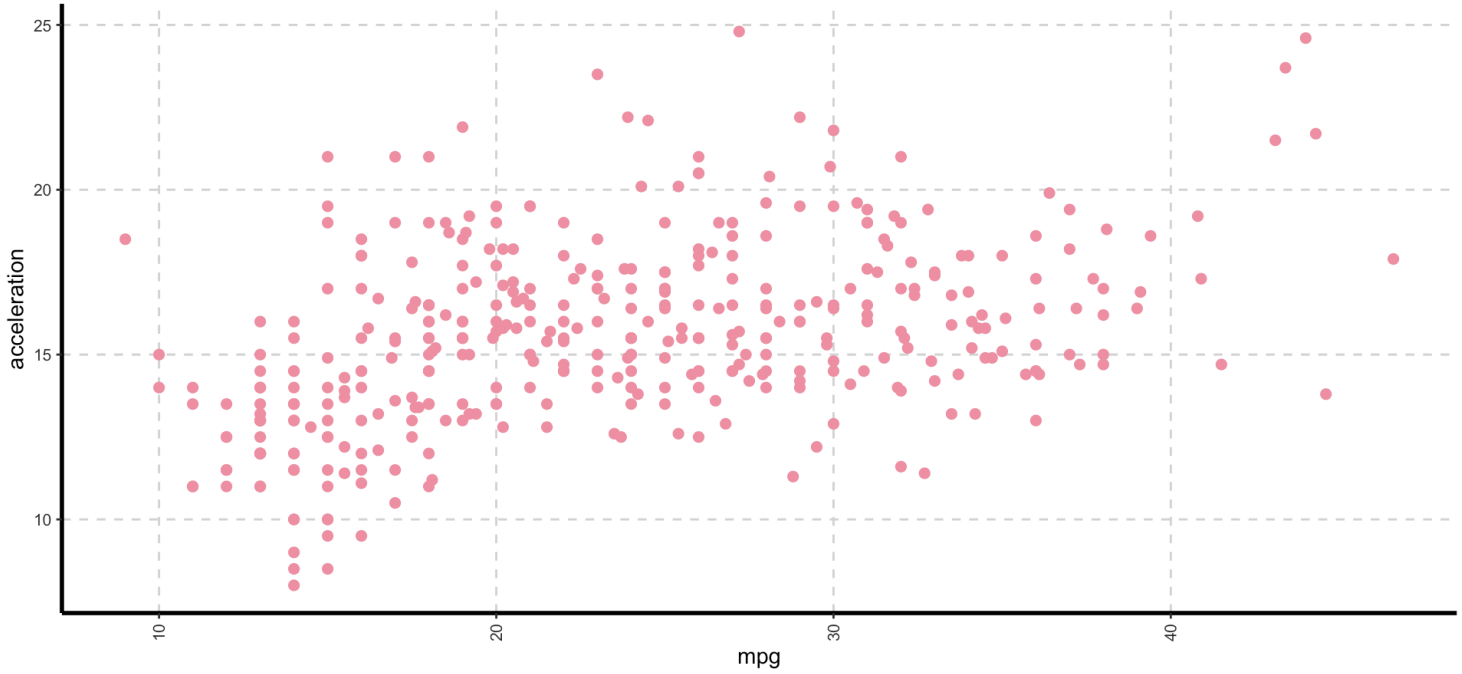
```
## $`0`
```


page 1 of 8
mpg vs cylinders
cylinders
8
7
6
5
4
3
10
20
30
40
mpg
mpg vs displacement
displacement
400
300
200
100
10
20
30
40
mpg

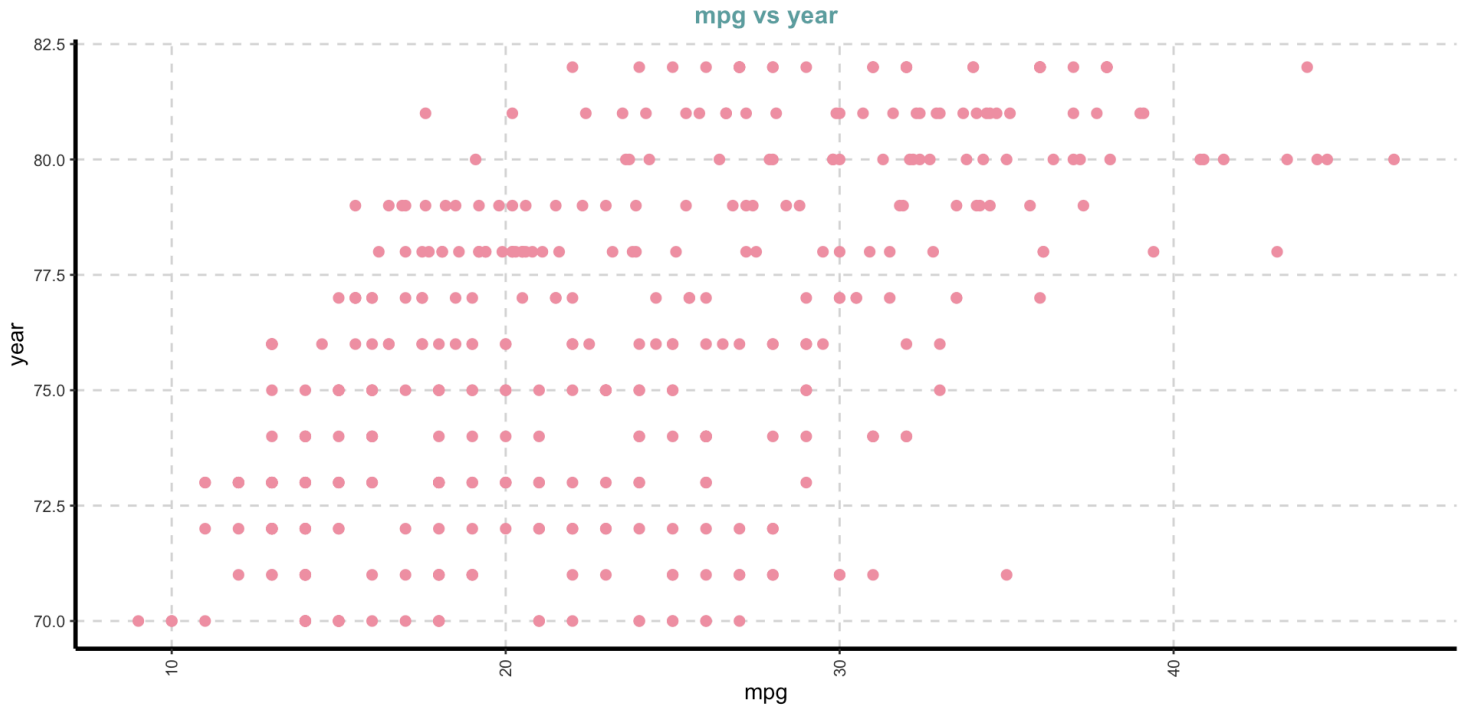
mpg vs weight

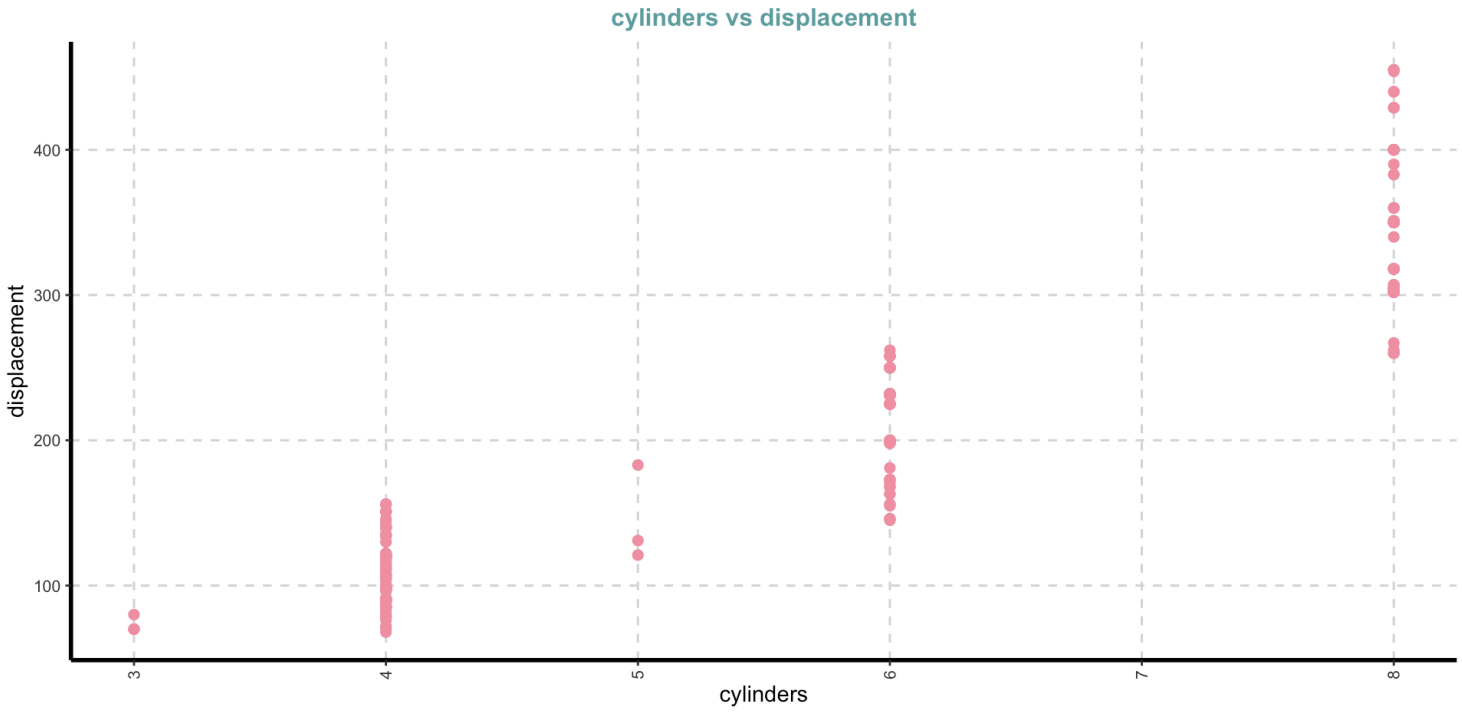


mpg vs acceleration



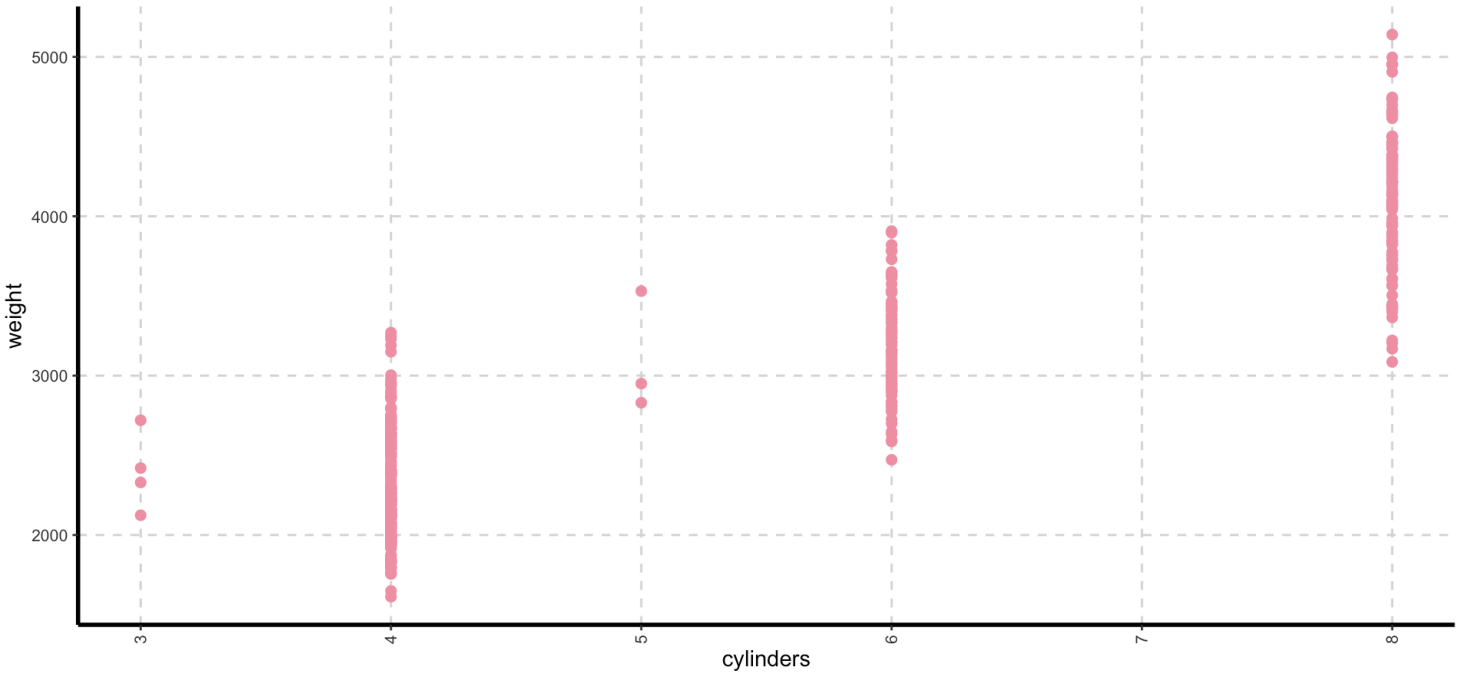
page 3 of 8


mpg vs year
year
mpg
82.5
80.0
77.5
75.0
72.5
70.0
10
20
30
40

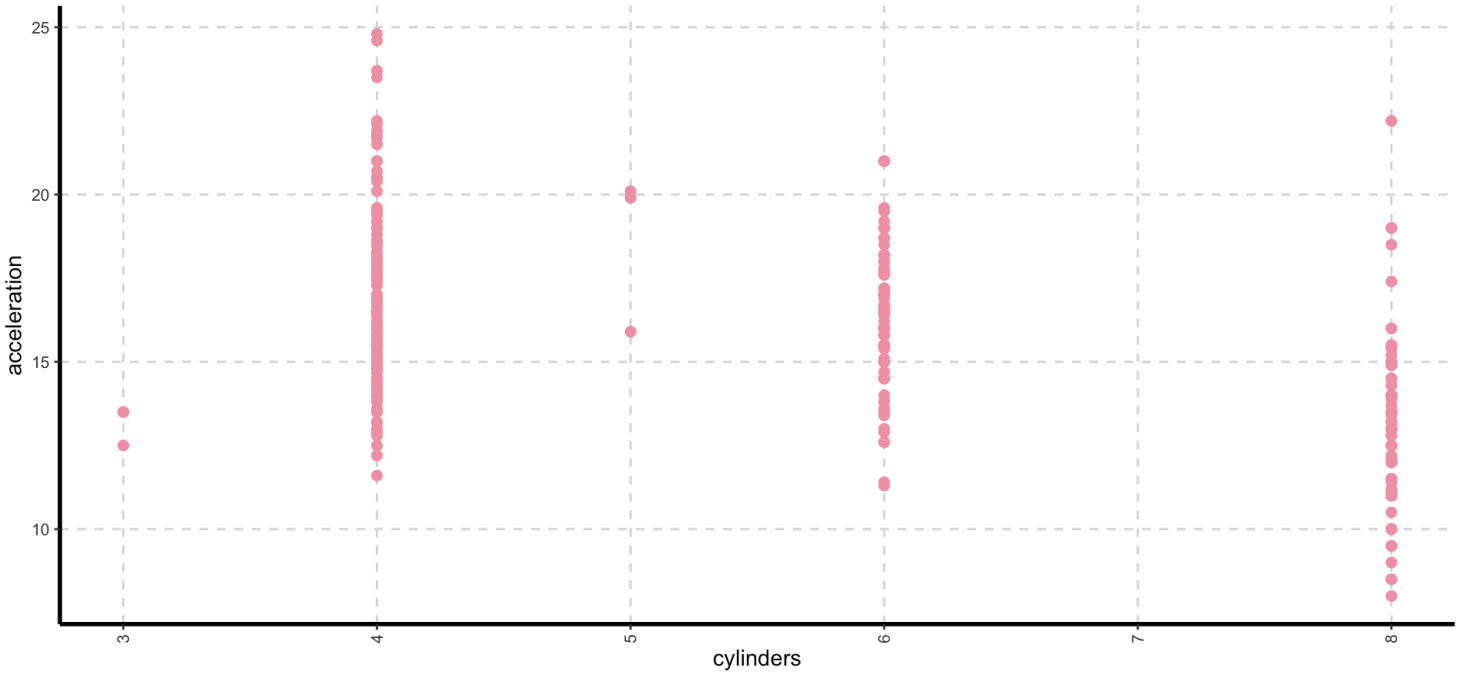

cylinders vs displacement
displacement
cylinders
400
300
200
100
3
4
5
6
7
8

page 4 of 8

cylinders vs weight

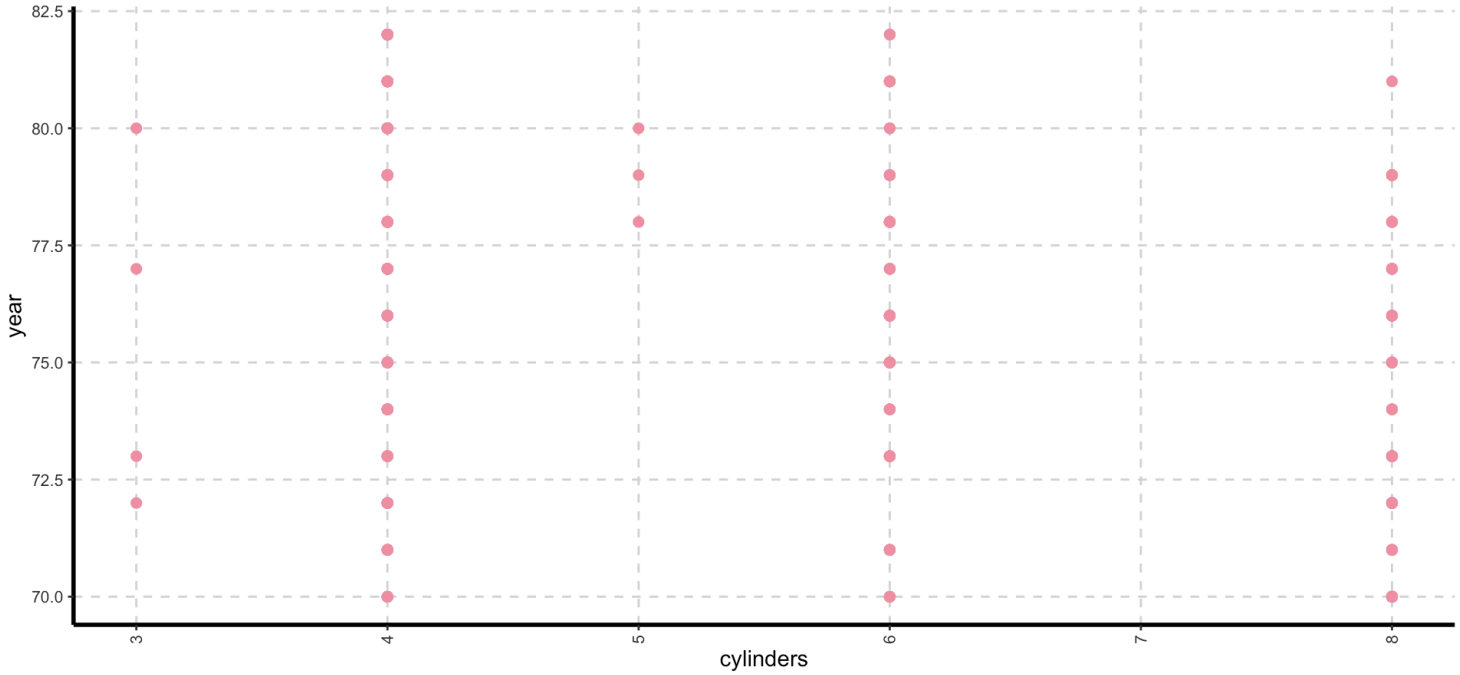


cylinders vs acceleration

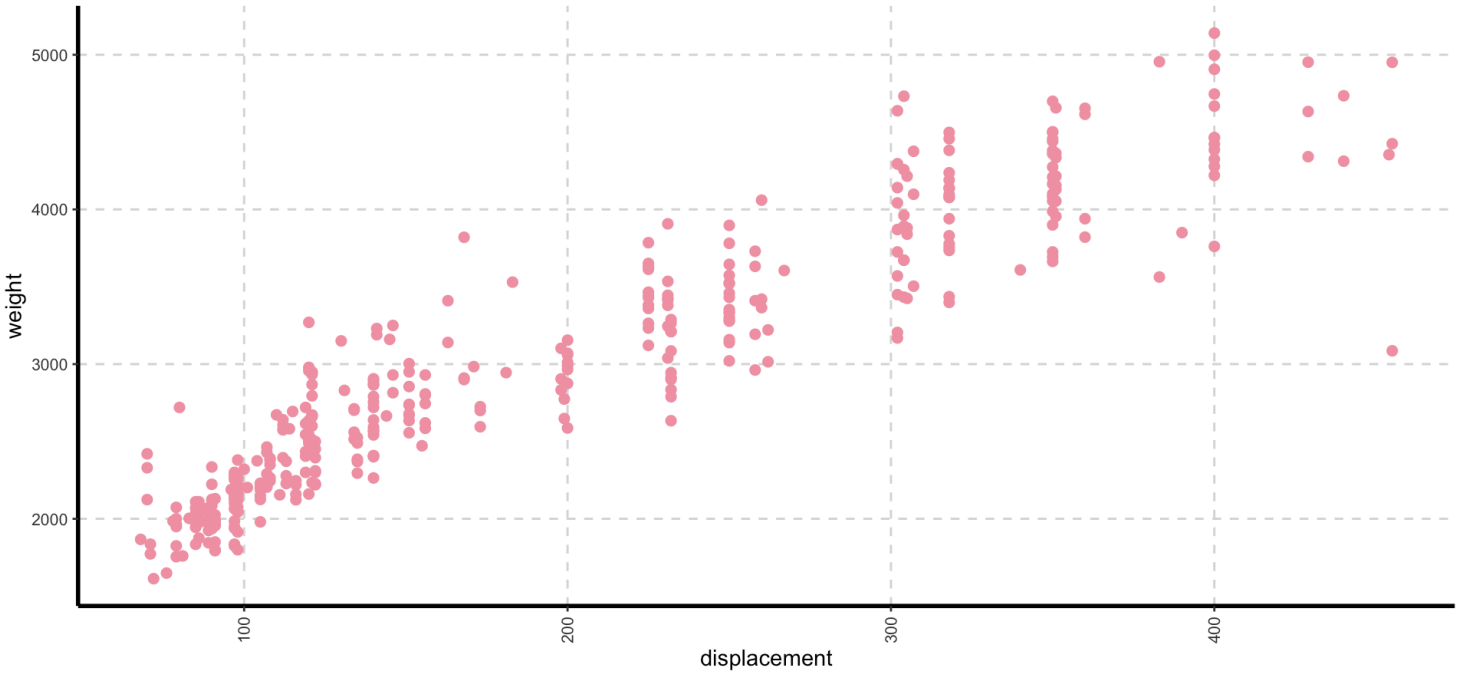


page 5 of 8

cylinders vs year



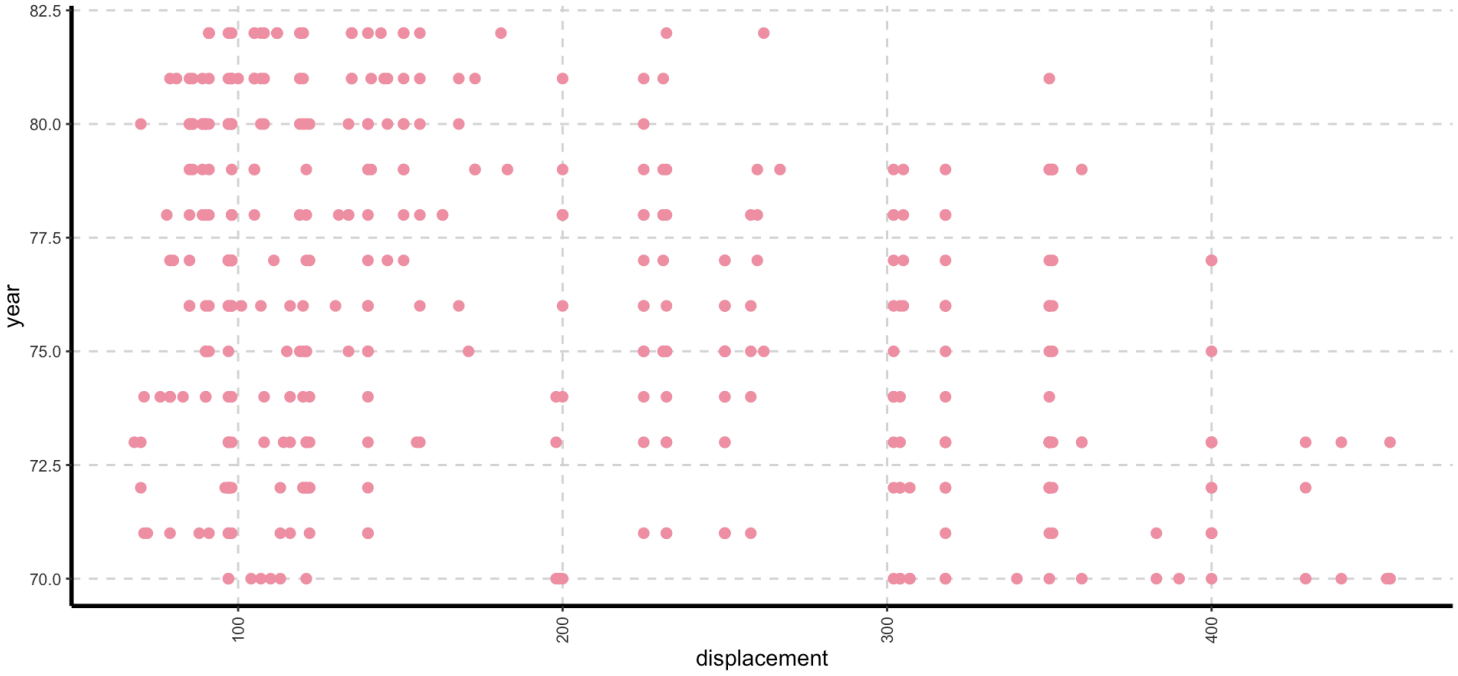
displacement vs weight



page 6 of 8

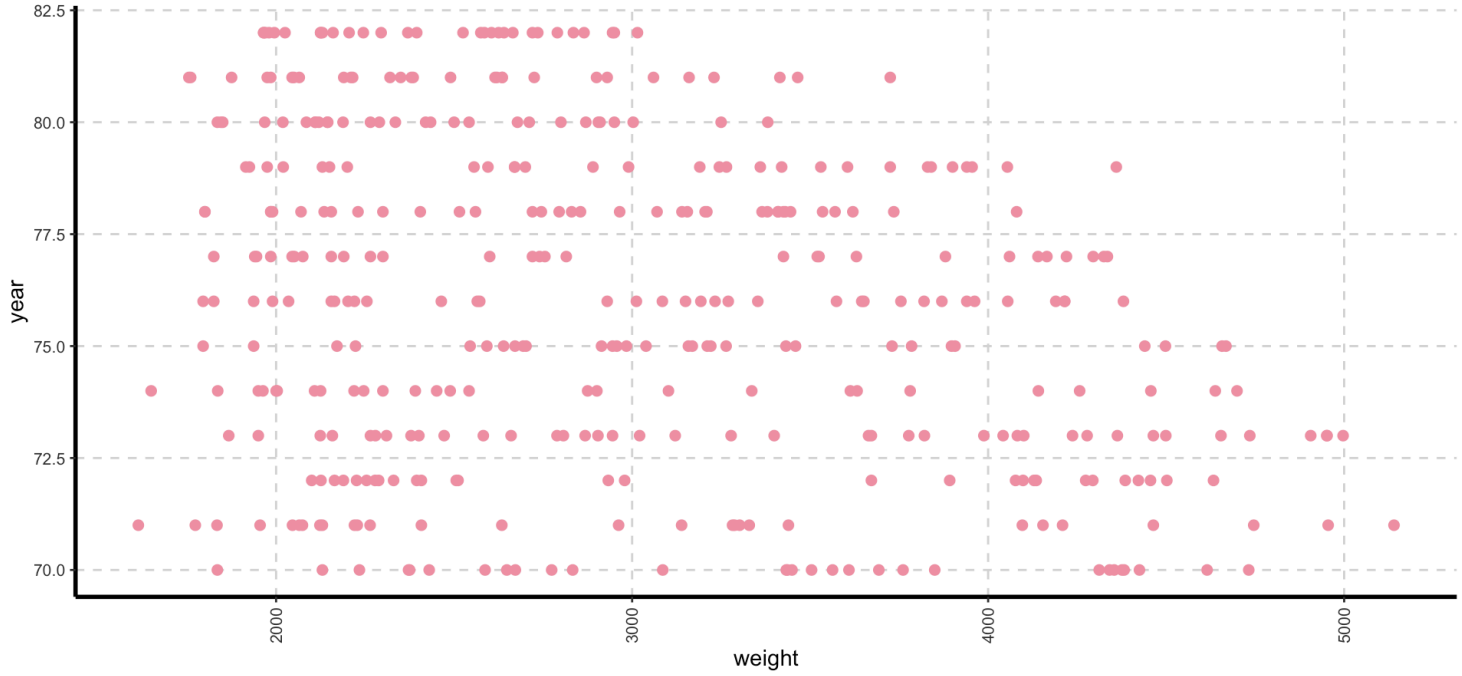
displacement vs acceleration

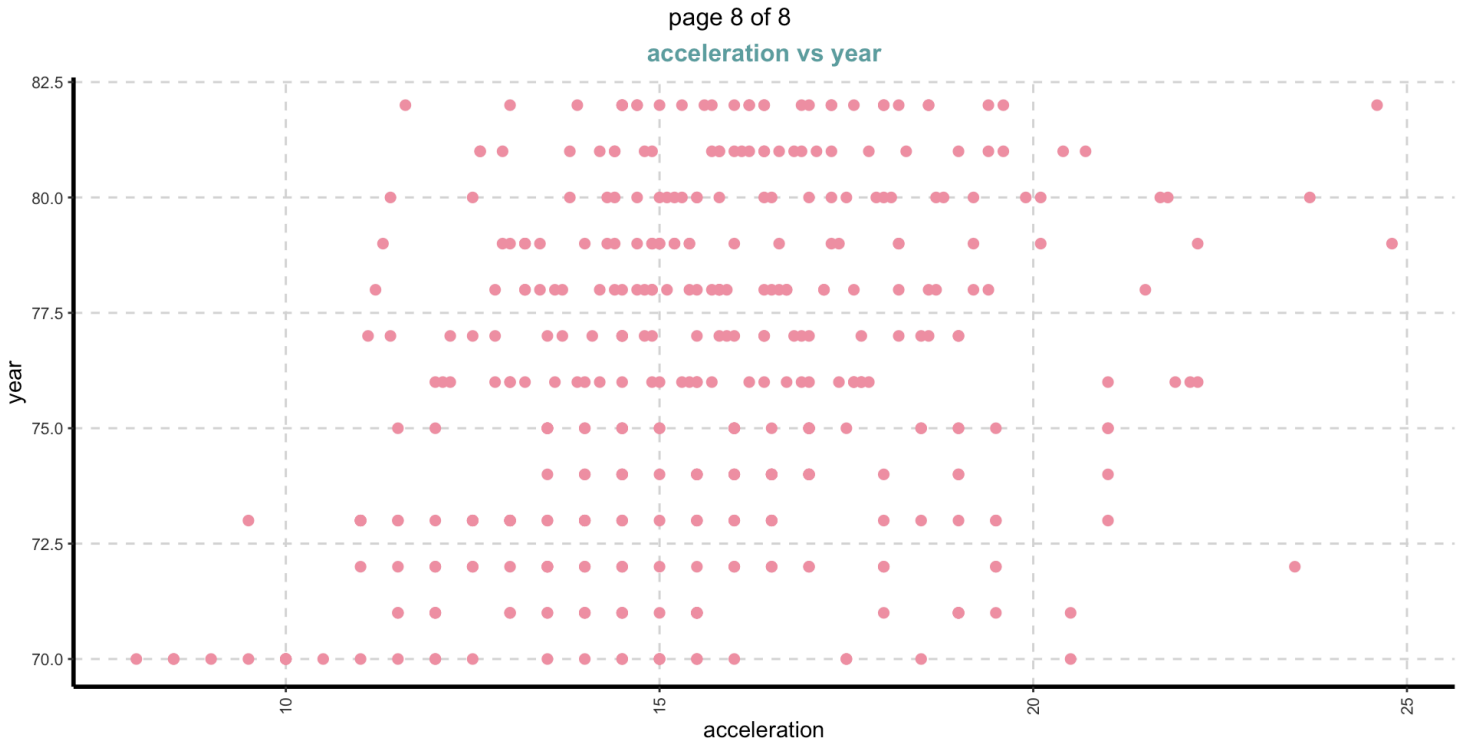
displacement vs year



weight vs acceleration

weight vs year



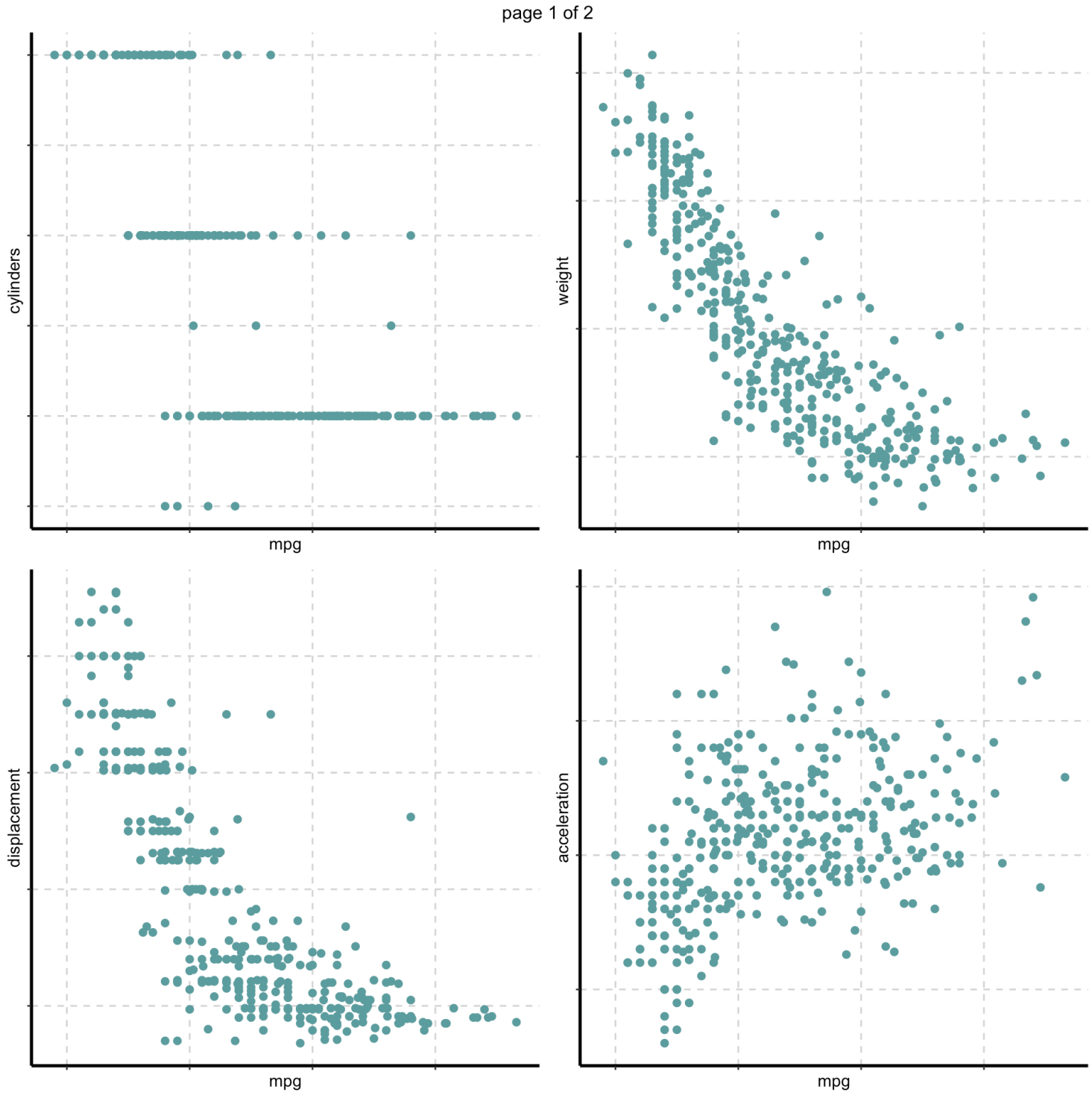


Correlation between dependent variable vs Independent variables

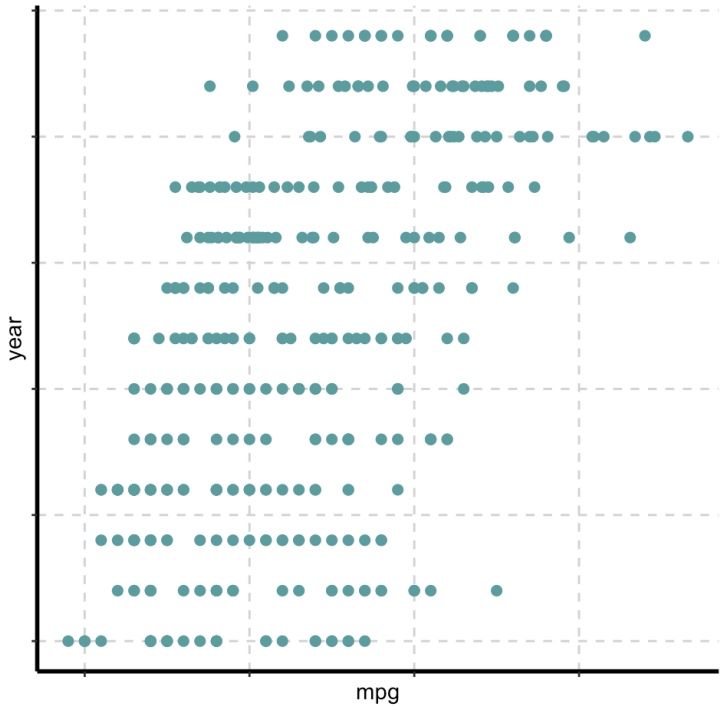
Dependent variable is **mpg** (continuous).

```
ExpNumViz(data,target=Target,nlim=5,fname=NULL,col=NULL,theme=theme,Page=c(2,2),sampl
e=sn)
```

$`0`


page 1 of 2
cylinders
mpg
weight
mpg
displacement
mpg
acceleration
mpg

page 2 of 2



** Correlation summary table

```
ExpNumStat(data,by="GA",gp=Target,MesofShape=2,Outlier=FALSE,round=2,dcast=T,val="cor
")
```

Stat	Vname	mpg
<chr>	<chr>	<dbl>

cor	acceleration	0.42
cor	displacement	-0.80
cor	mpg	1.00
cor	weight	-0.83
cor	year	0.58

5 rows

4. Summary of categorical variables

Summary of categorical variables

- frequency for all categorical independent variables

```
ExpCTable(data,margin=1,clim=10,nlim=5,round=2,per=T)
```

Variable <chr>	**Valid** <chr>	**Frequency** <dbl>	**Percent** <dbl>	**CumPercent** <dbl>
cylinders	3	4	1.01	1.01
cylinders	4	203	51.13	52.14
cylinders	5	3	0.76	52.90
cylinders	6	84	21.16	74.06
cylinders	8	103	25.94	100.00
cylinders	TOTAL	397	*NA*	*NA*
origin	1	248	62.47	62.47
origin	2	70	17.63	80.10
origin	3	79	19.90	100.00
origin	TOTAL	397	*NA*	*NA*

1-10 of 10 rows

- frequency for all categorical independent variables by descretized **mpg**

```
##bin=4, descretized 4 categories based on quantiles
ExpCTable(data,Target=Target,margin=1,clim=10,nlim=5,round=2,bin=4,per=T)
```

VARIABLE <chr>	CATEG... <chr>	Nu... <chr>	mpg:(8.96,18.4] <dbl>	mpg:(18.4,27.8] <dbl>	mpg:(27.8,37.2] <dbl>	mpg:(
cylinders	3	nn	1.00	3.00	0	
cylinders	4	nn	1.00	88.00	96	
cylinders	5	nn	0.00	2.00	1	
cylinders	6	nn	32.00	48.00	3	
cylinders	8	nn	93.00	10.00	0	
cylinders	TOTAL	nn	127.00	151.00	100	
cylinders	3	%	0.79	1.99	0	
cylinders	4	%	0.79	58.28	96	
cylinders	5	%	0.00	1.32	1	
cylinders	6	%	25.20	31.79	3	

1-10 of 20 rows Previous **1** 2 Next

5. Distributions of Categorical variables

Graphical representation of all Categorical variables

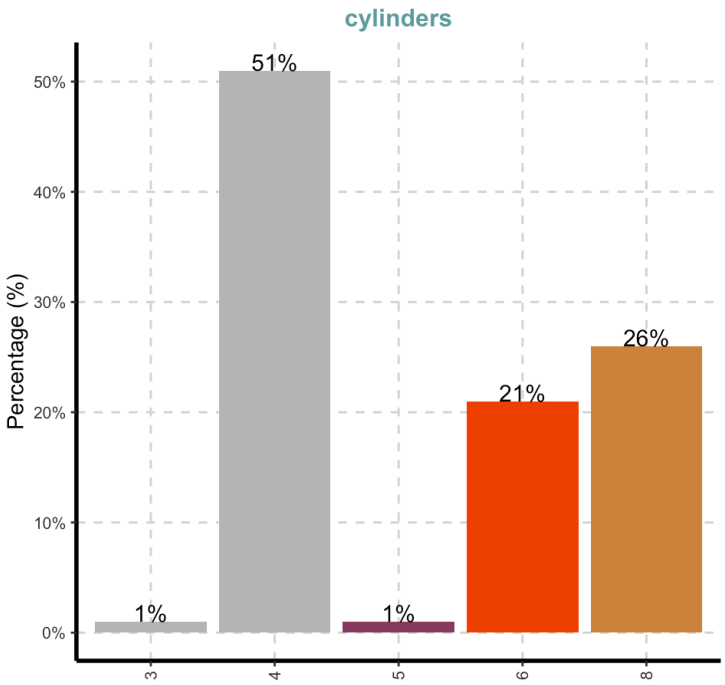
- Bar plot (Univariate)

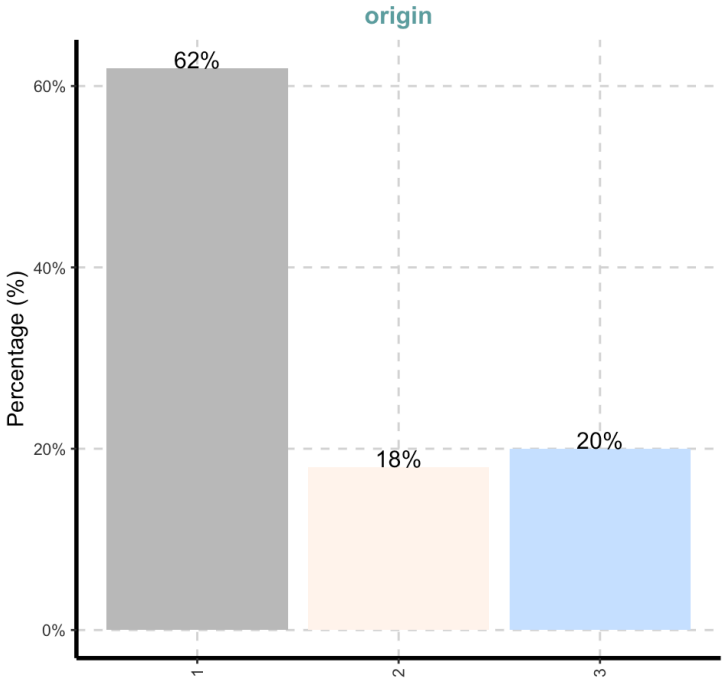
Bar plot with vertical or horizontal bars for all categorical variables

```
ExpCatViz(data,clim=10,margin=2,theme=theme,Page = c(2,2),sample=sc)
```

```
## $`0`
```

page 1 of 1

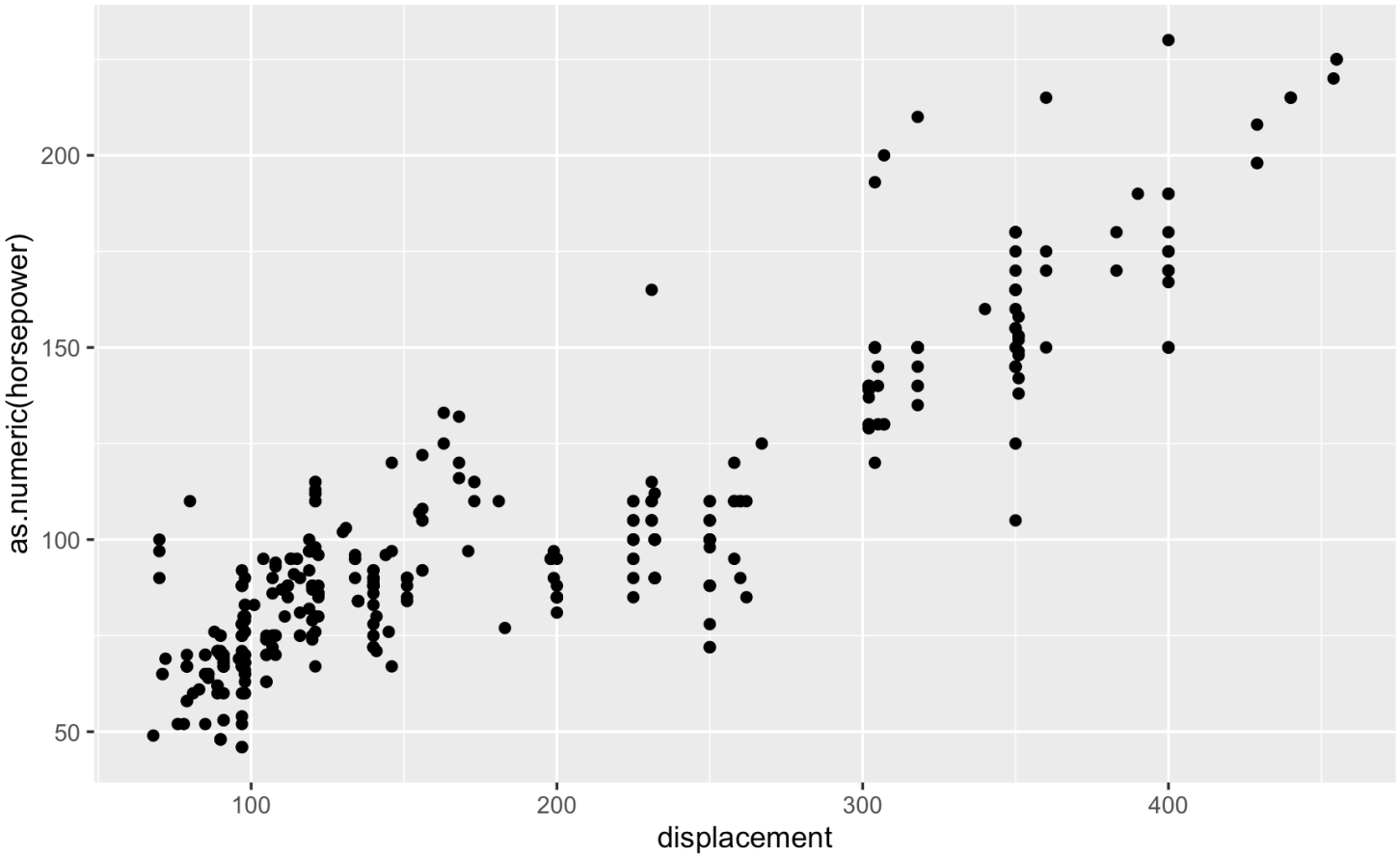

cylinders
Percentage (%)
50%
40%
30%
20%
10%
0%
51%
26%
21%
1%
1%
3
4
5
6
8


origin
Percentage (%)
60%
40%
20%
0%
62%
18%
20%
1
2
3

Hide

```
auto %>%
  ggplot() +
  geom_point(aes(x = displacement, y = as.numeric(horsepower)))
```

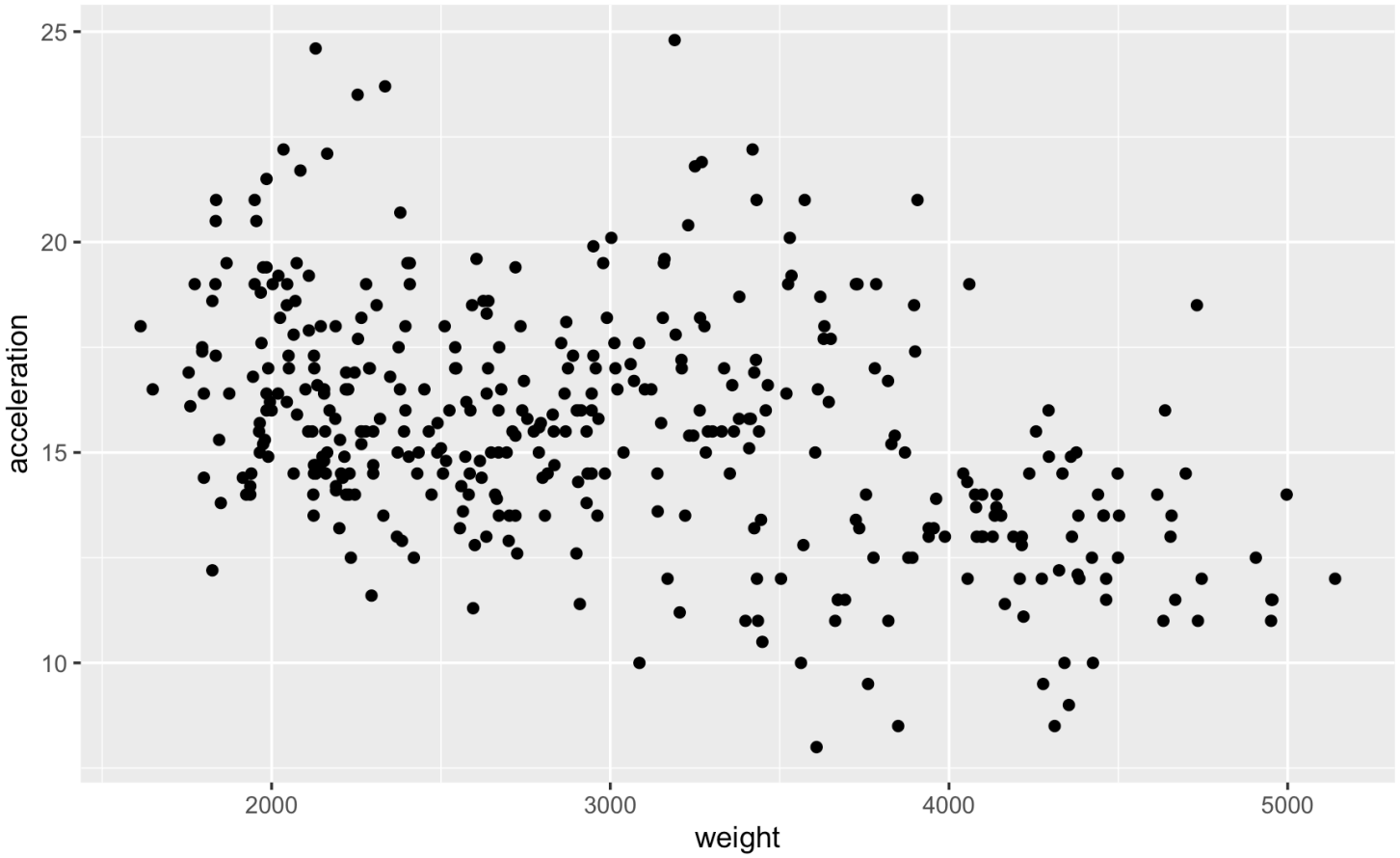
```
Warning in FUN(X[[i]], ...) : NAs introduced by coercion
Warning in FUN(X[[i]], ...) : NAs introduced by coercion
Warning: Removed 5 rows containing missing values (geom_point).
```



Wow. Displacement generally correlates with horsepower. Who could have predicted this?

Hide

```
auto %>%
  ggplot() +
  geom_point(aes(x = weight, y = acceleration))
```



Hide

```
bigCarGoSlow <- lm(acceleration ~ weight + displacement, data = auto)
bigCarGoSlow
```

```

Call:
lm(formula = acceleration ~ weight + displacement, data = auto)

Coefficients:
 (Intercept)        weight  displacement
   15.002648      0.002214     -0.031115
```

Hide

```
summary(bigCarGoSlow)
```

```
Call:
lm(formula = acceleration ~ weight + displacement, data = auto)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6813 -1.5377 -0.1879  1.2041  7.9007

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.0026477  0.5959068  25.176  < 2e-16 ***
weight        0.0022135  0.0003645   6.072 2.97e-09 ***
displacement -0.0311147  0.0029612 -10.507  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.212 on 394 degrees of freedom
Multiple R-squared:  0.3563,    Adjusted R-squared:  0.3531
F-statistic: 109.1 on 2 and 394 DF,  p-value: < 2.2e-16
```

Horsepower should really be converted to numeric but I don’t care. Big car go slow is probably correct unless it’s an electric Hummer.

9f

> Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg` ? Justify your answer.

Hide

```
auto$horsepower <- as.numeric(auto$horsepower)
```

```
Warning: NAs introduced by coercion
```

Hide

```
mpg <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year
+ origin, data = auto)
mpg
```

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, data = auto)

Coefficients:
 (Intercept)     cylinders  displacement    horsepower        weight  acceleration
year        origin
  -17.218435     -0.493376      0.019896     -0.016951     -0.006474      0.080576
0.750773      1.426140
```

Hide

```
summary(mpg)
```

```
Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, data = auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Based on p-values, `displacement`, `weight`, `year`, and `origin` are all good predictors for `mpg`.

10

> This exercise involves the Boston housing data set.

10a

To begin, load in the `Boston` data set. The `Boston` data set is part of the `ISLR2` *library*. > > `library(ISLR2)` > > Now the data set is contained in the object Boston. > Boston > > Read about the data set: > `?Boston` > > How many rows are in this data set? How many columns? What do the rows and columns represent?

Hide

```
library(ISLR2)
Boston
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5

1-10 of 506 rows | 1-10 of 13 columns Previous **1** 2 3 4 5 6 … 51 Next

Hide

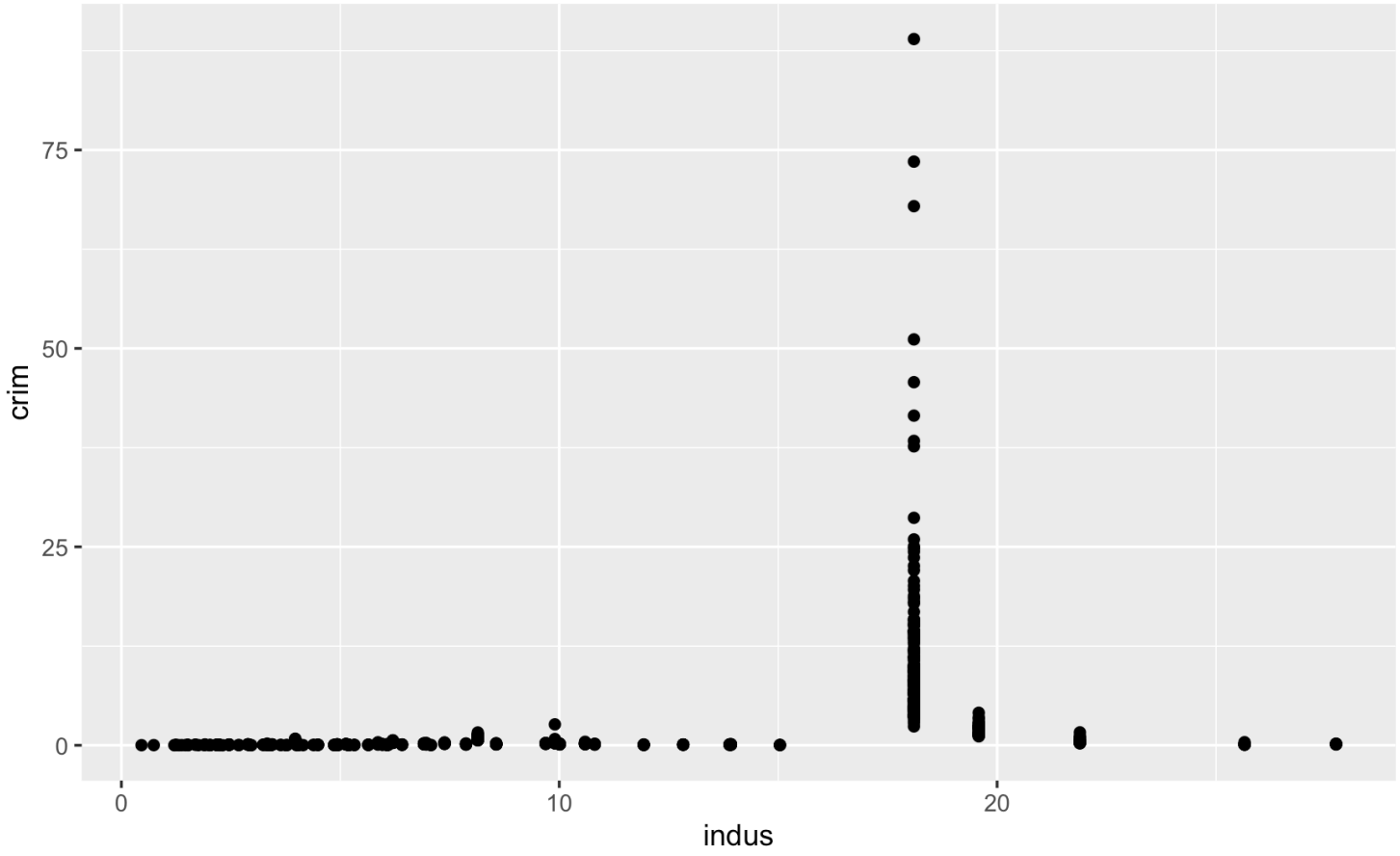
```
?Boston
```

10b

Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

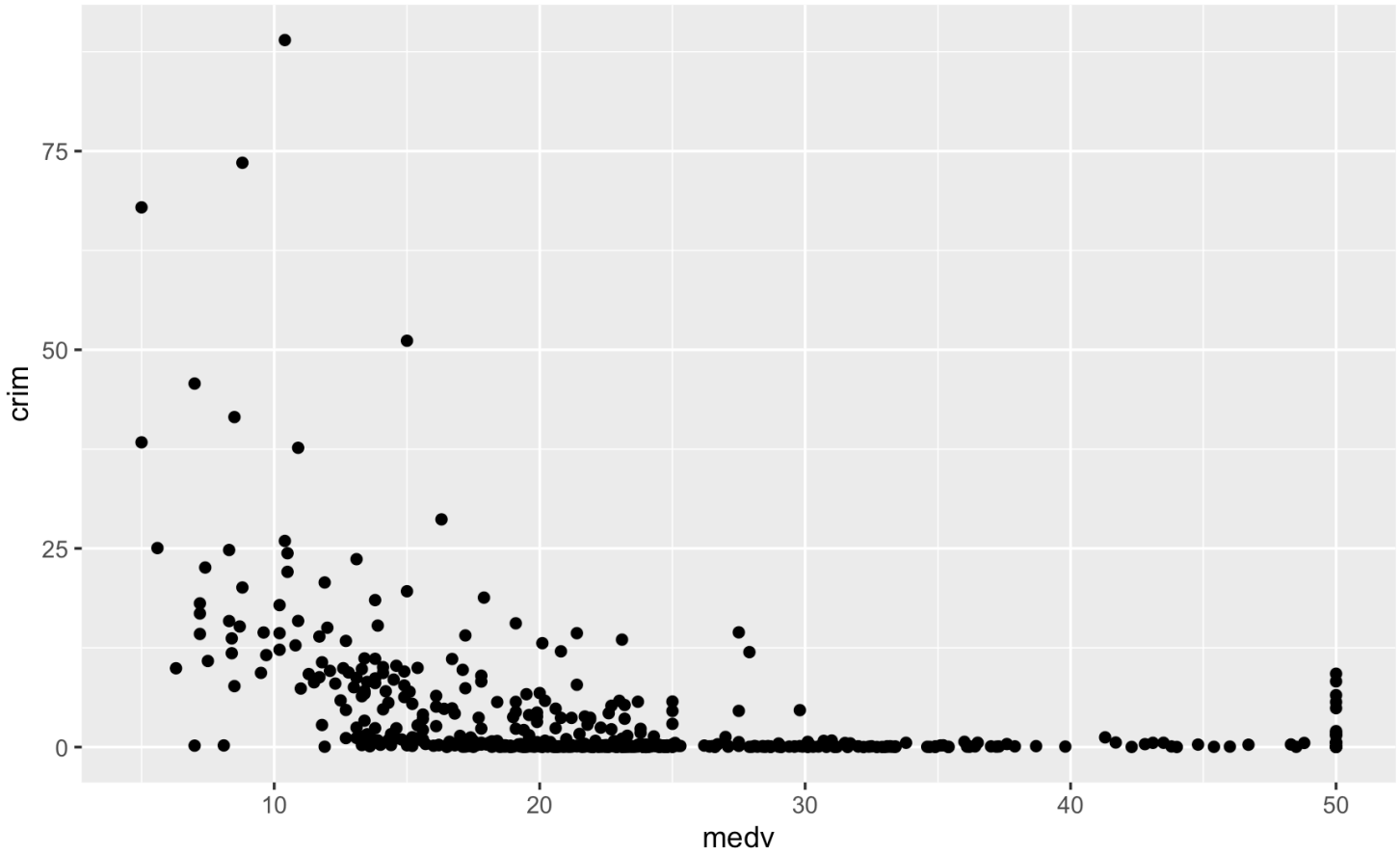
Hide

```
# Predict crime rate based on industry
Boston %>%
  ggplot() +
  geom_point(aes(x = indus, y = crim))
```



Hide

```
# Predict crime rate based on median home value
Boston %>%
  ggplot() +
  geom_point(aes(x = medv, y = crim))
```



Crime rates are low based on my selection criteria or maybe some places just have too much crime. I just looked at the next question. I hate this.

10c

> Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Hide

```
crim <- lm(crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + ls
tat + medv, data = Boston)
crim
```

```
Call:
lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
    rad + tax + ptratio + lstat + medv, data = Boston)

Coefficients:
(Intercept)           zn        indus         chas          nox           rm
age          dis
 13.7783938    0.0457100   -0.0583501   -0.8253776   -9.9575865    0.6289107   -0.000
8483   -1.0122467
        rad          tax      ptratio        lstat         medv
  0.6124653   -0.0037756   -0.3040728    0.1388006   -0.2200564
```

Hide

```
summary(crim)
```

```
Call:
lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
    rad + tax + ptratio + lstat + medv, data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-8.534 -2.248 -0.348  1.087 73.923

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.7783938  7.0818258   1.946 0.052271 .
zn           0.0457100  0.0187903   2.433 0.015344 *
indus       -0.0583501  0.0836351  -0.698 0.485709
chas        -0.8253776  1.1833963  -0.697 0.485841
nox         -9.9575865  5.2898242  -1.882 0.060370 .
rm           0.6289107  0.6070924   1.036 0.300738
age         -0.0008483  0.0179482  -0.047 0.962323
dis         -1.0122467  0.2824676  -3.584 0.000373 ***
rad          0.6124653  0.0875358   6.997 8.59e-12 ***
tax         -0.0037756  0.0051723  -0.730 0.465757
ptratio     -0.3040728  0.1863598  -1.632 0.103393
lstat        0.1388006  0.0757213   1.833 0.067398 .
medv        -0.2200564  0.0598240  -3.678 0.000261 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.46 on 493 degrees of freedom
Multiple R-squared:  0.4493,    Adjusted R-squared:  0.4359
F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

`zn`, `dis`, `rad`, and `medv` are good predictors. `zn` and `rad` correlate with an increase in crime while `dis` and `medv` correlate with a decrease in crime.

10d

Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Hide

```
# DataExplorer::create_report(data = Boston,
#                             output_file = "10d.html",
#                             report_title = "Exploratory Data Analysis Report for
#                             the Boston Dataset with a Focus on the Per Capita
#                             Crime Rate",
#                             y = "crim")
# rmarkdown::pandoc_convert("10d.html", to = "pdf", from = "html", output = "10d.pdf"
)
```

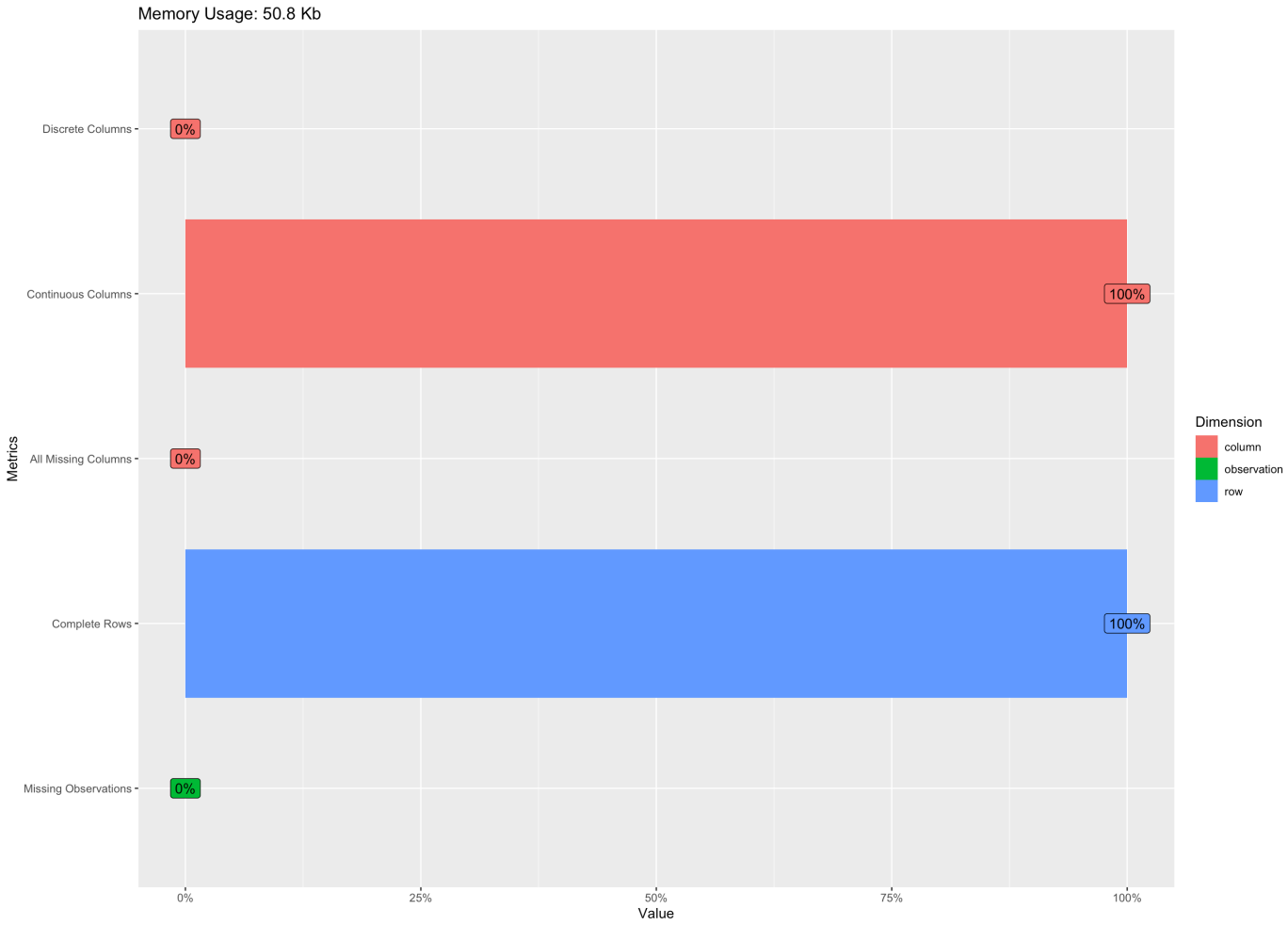
Exploratory Data Analysis Report for the Boston Dataset with a Focus on the Per Capita Crime Rate

Basic Statistics

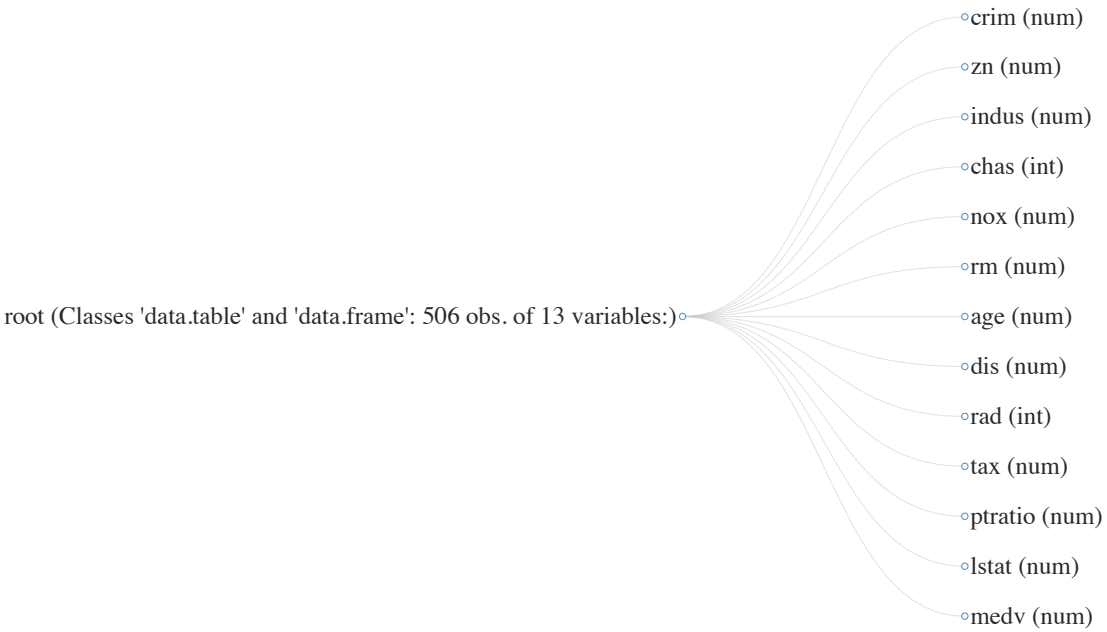
Raw Counts

Name	Value
Rows	506
Columns	13
Discrete columns	0
Continuous columns	13
All missing columns	0
Missing observations	0
Complete Rows	506
Total observations	6,578
Memory allocation	50.8 Kb

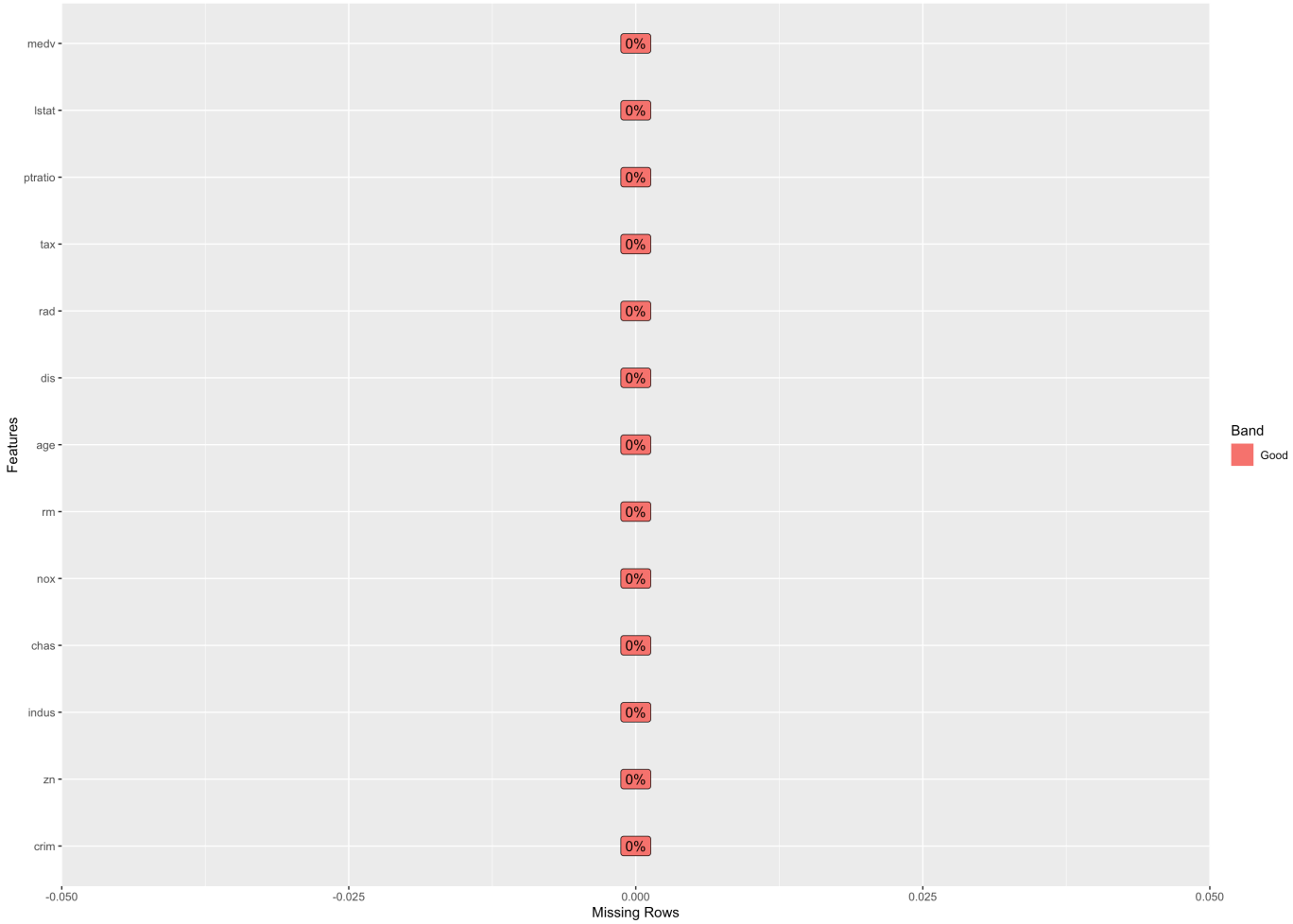
Percentages



Data Structure

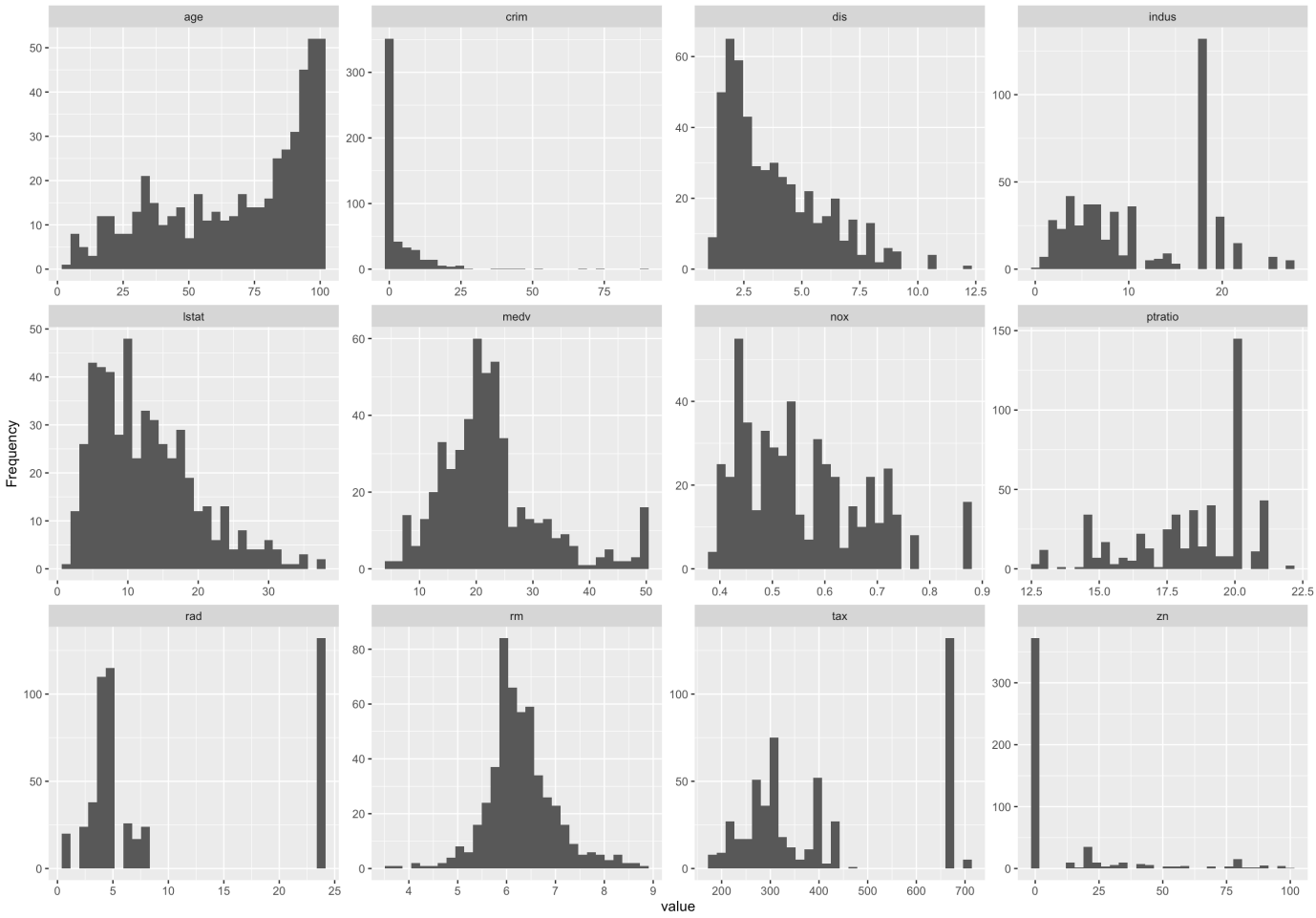


Missing Data Profile

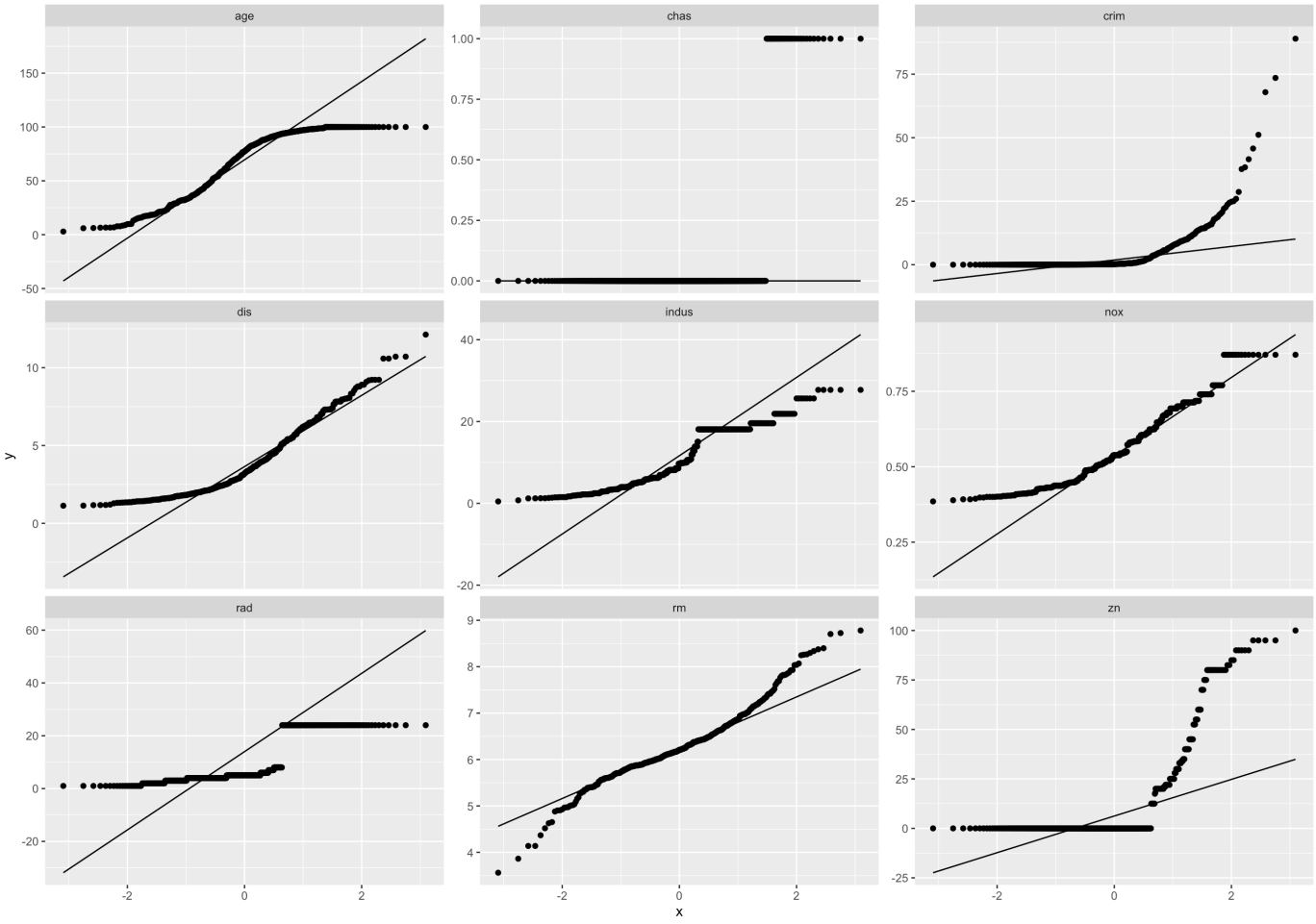


Univariate Distribution

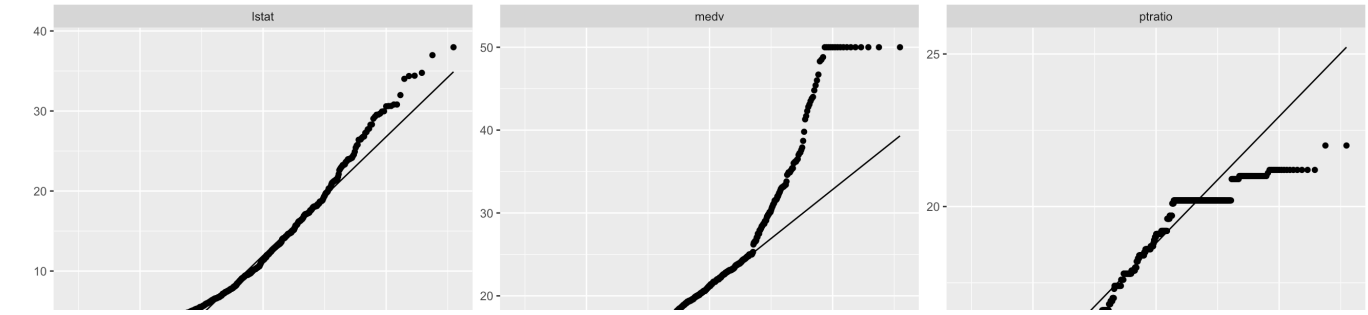
Histogram

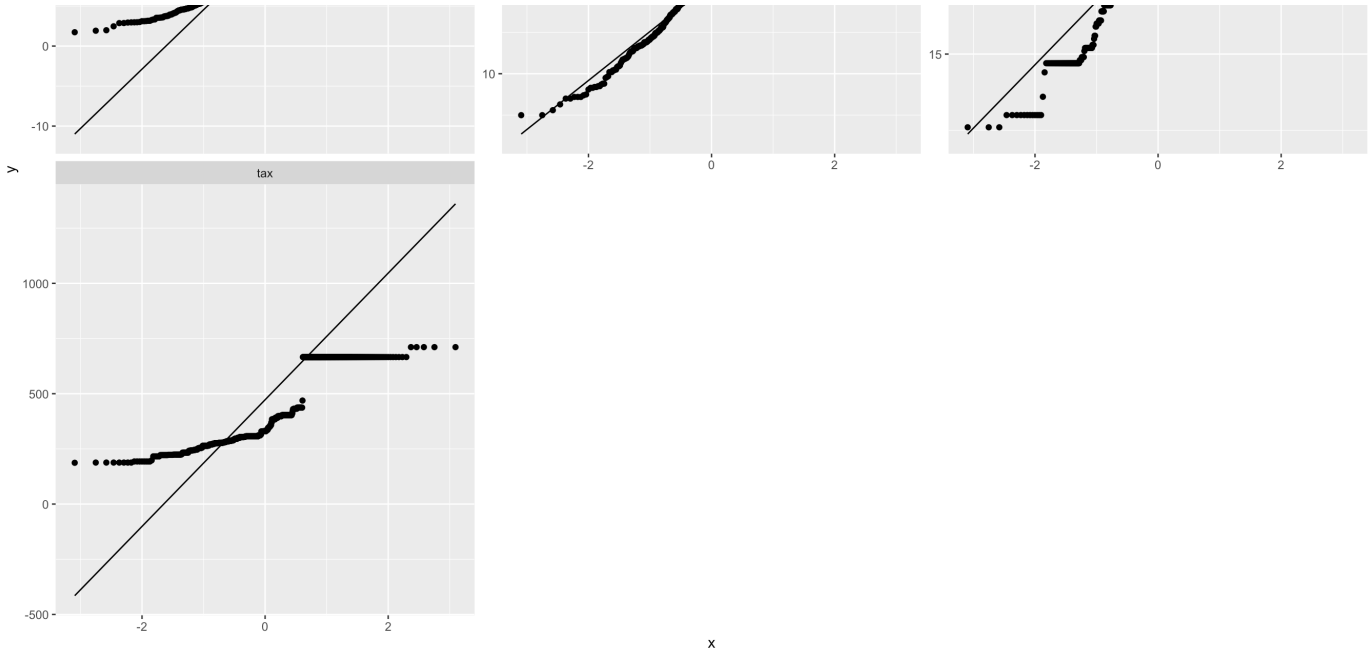

age
crim
dis
indus
lstat
medv
nox
ptratio
rad
rm
tax
zn
Frequency
value

QQ Plot



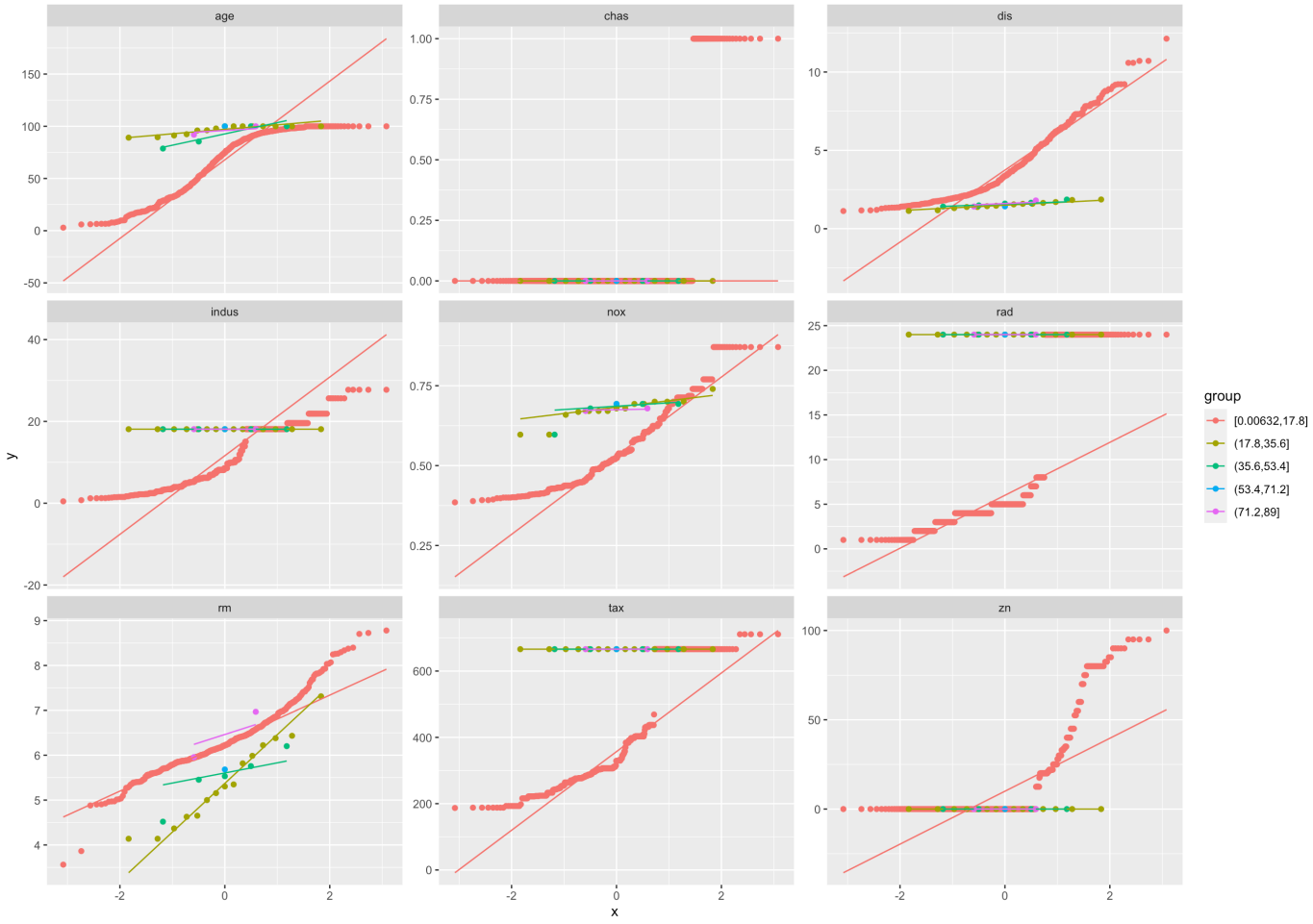
Page 1



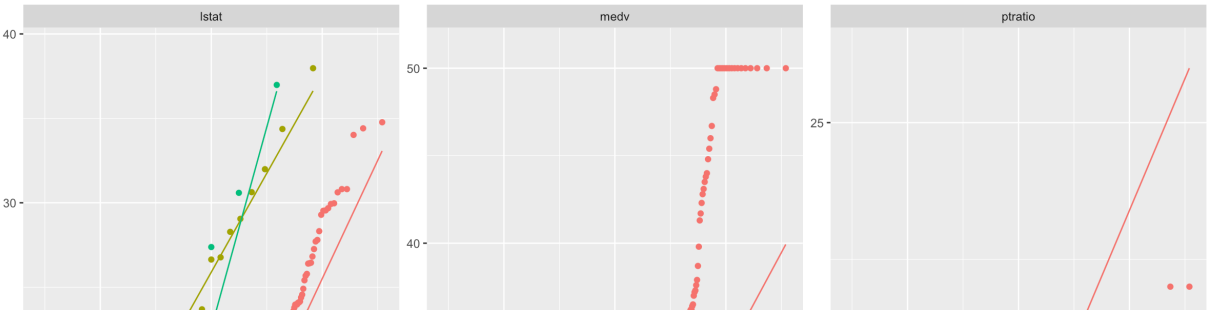


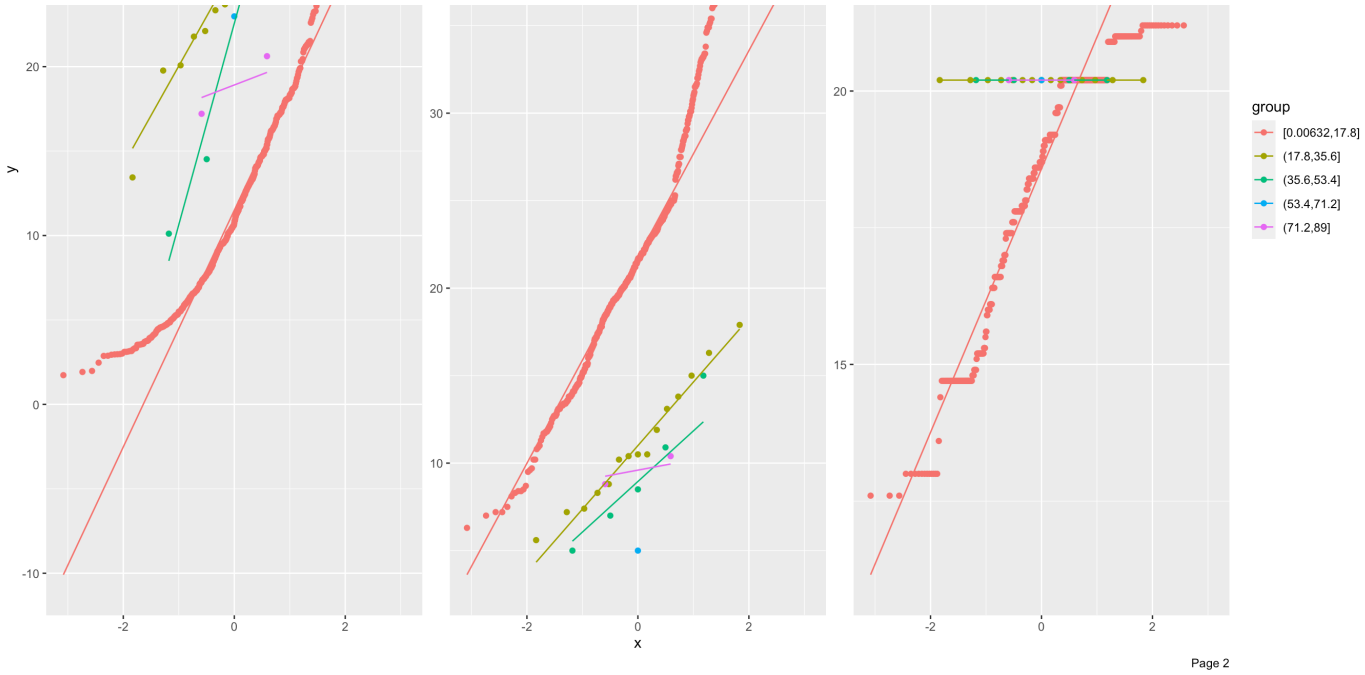
Page 2

QQ Plot (by crim)

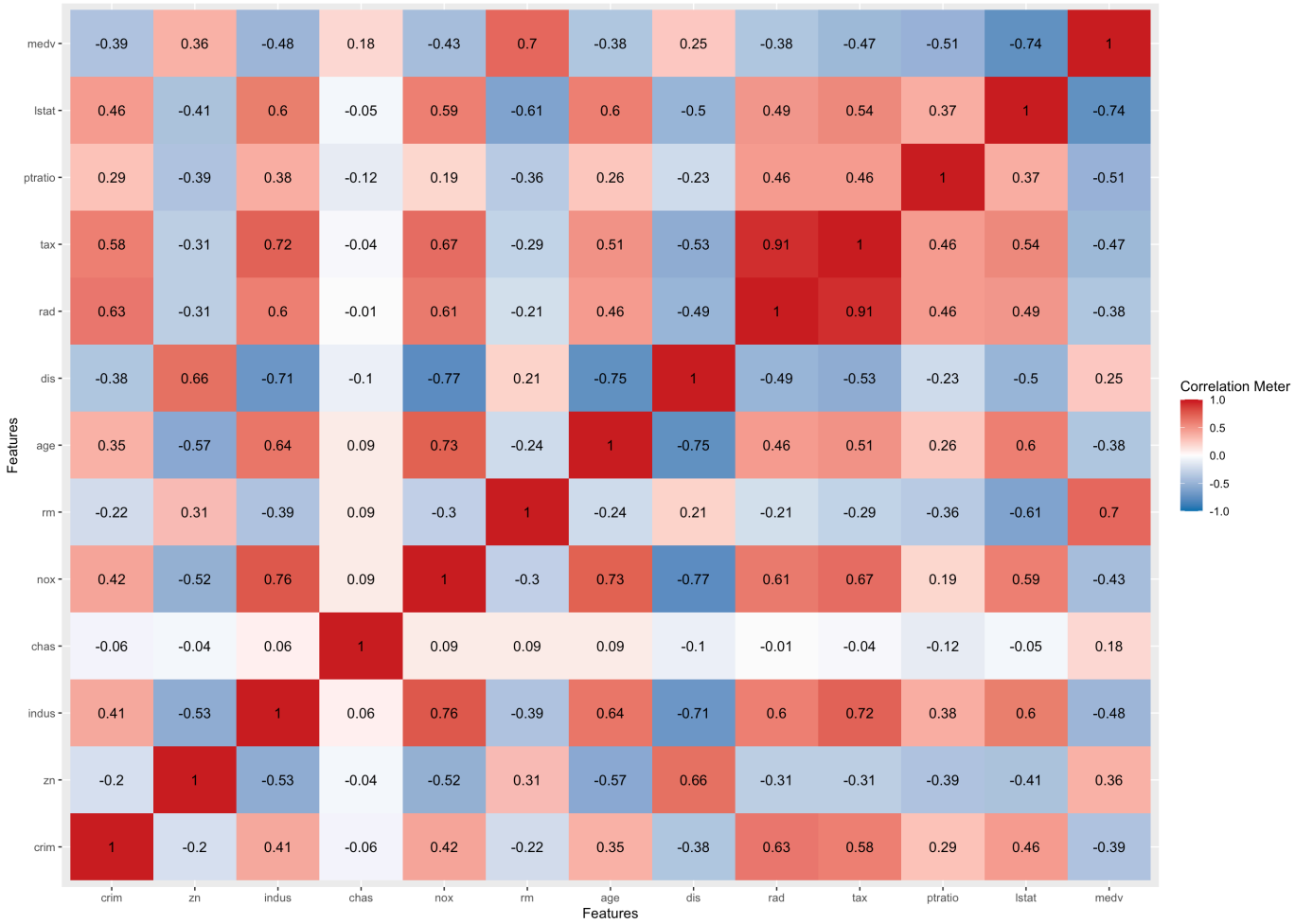

age
chas
dis
indus
nox
rad
rm
tax
zn
y
x
group
[0.00632,17.8]
(17.8,35.6]
(35.6,53.4]
(53.4,71.2]
(71.2,89]

Page 1

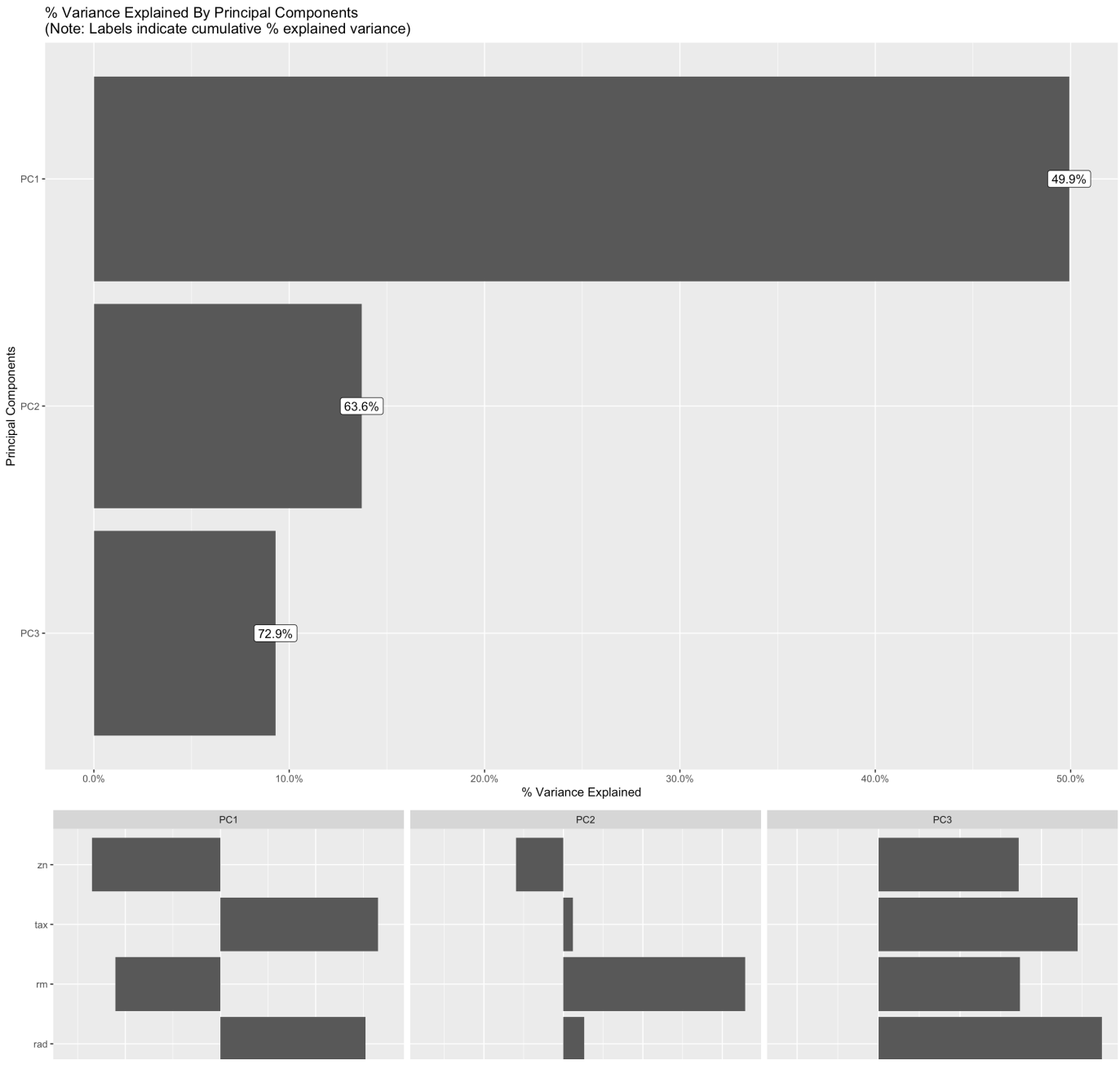

lstat
medv
ptratio

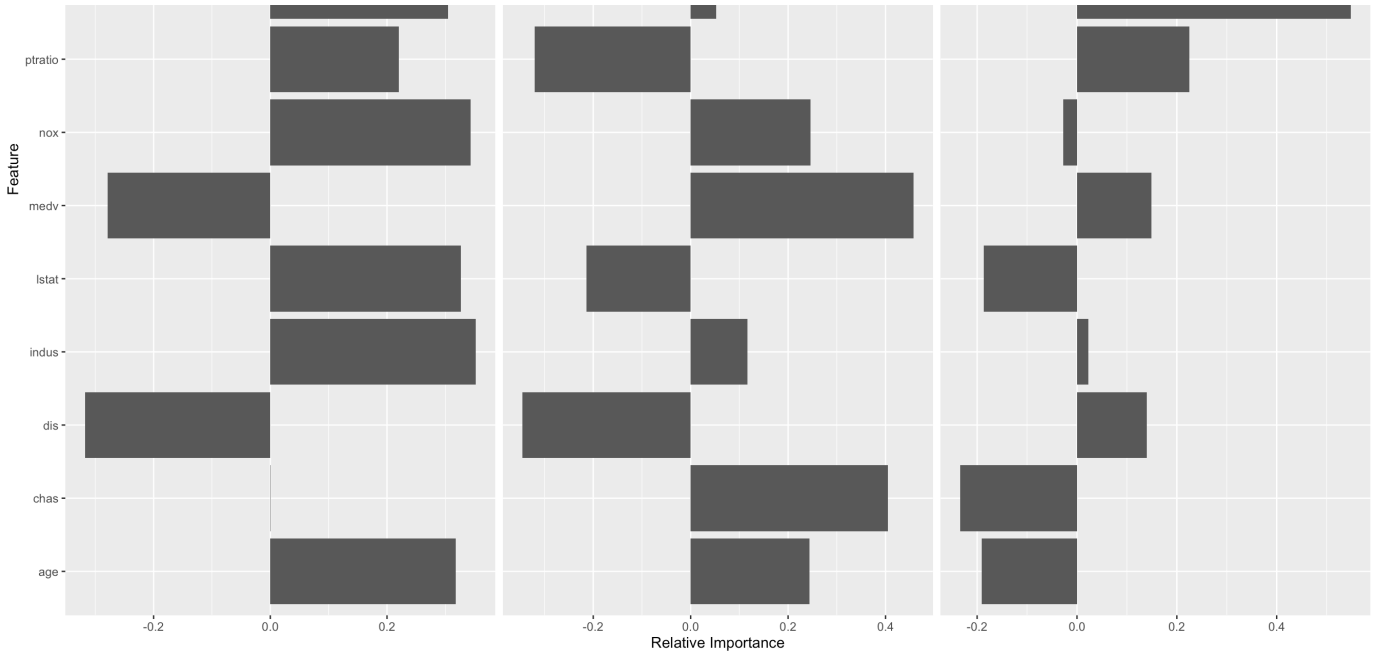


Correlation Analysis



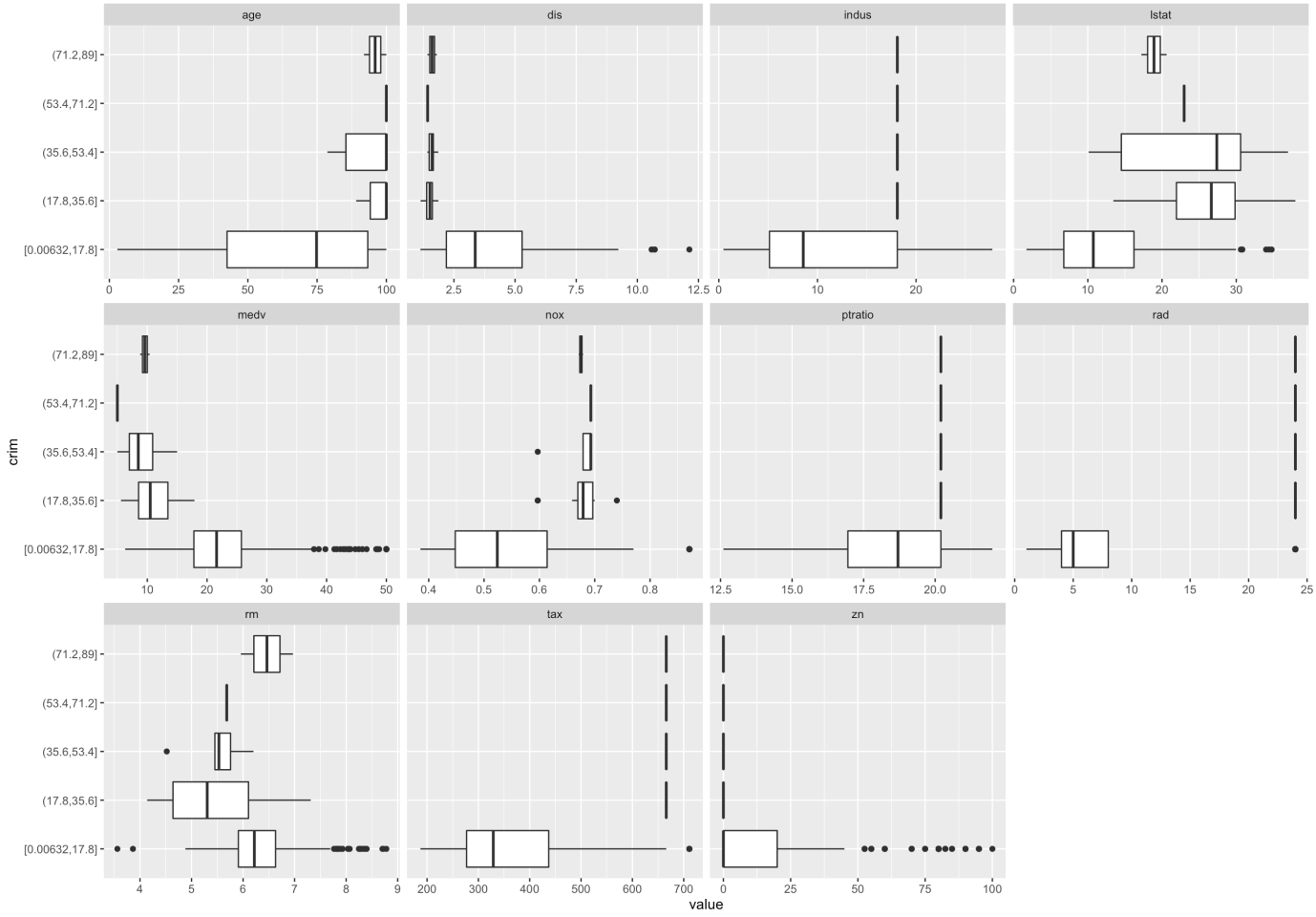
Principal Component Analysis


% Variance Explained By Principal Components
(Note: Labels indicate cumulative % explained variance)
Principal Components
PC1
PC2
PC3
49.9%
63.6%
72.9%
0.0%
10.0%
20.0%
30.0%
40.0%
50.0%
% Variance Explained
PC1
PC2
PC3
zn
tax
rm
rad

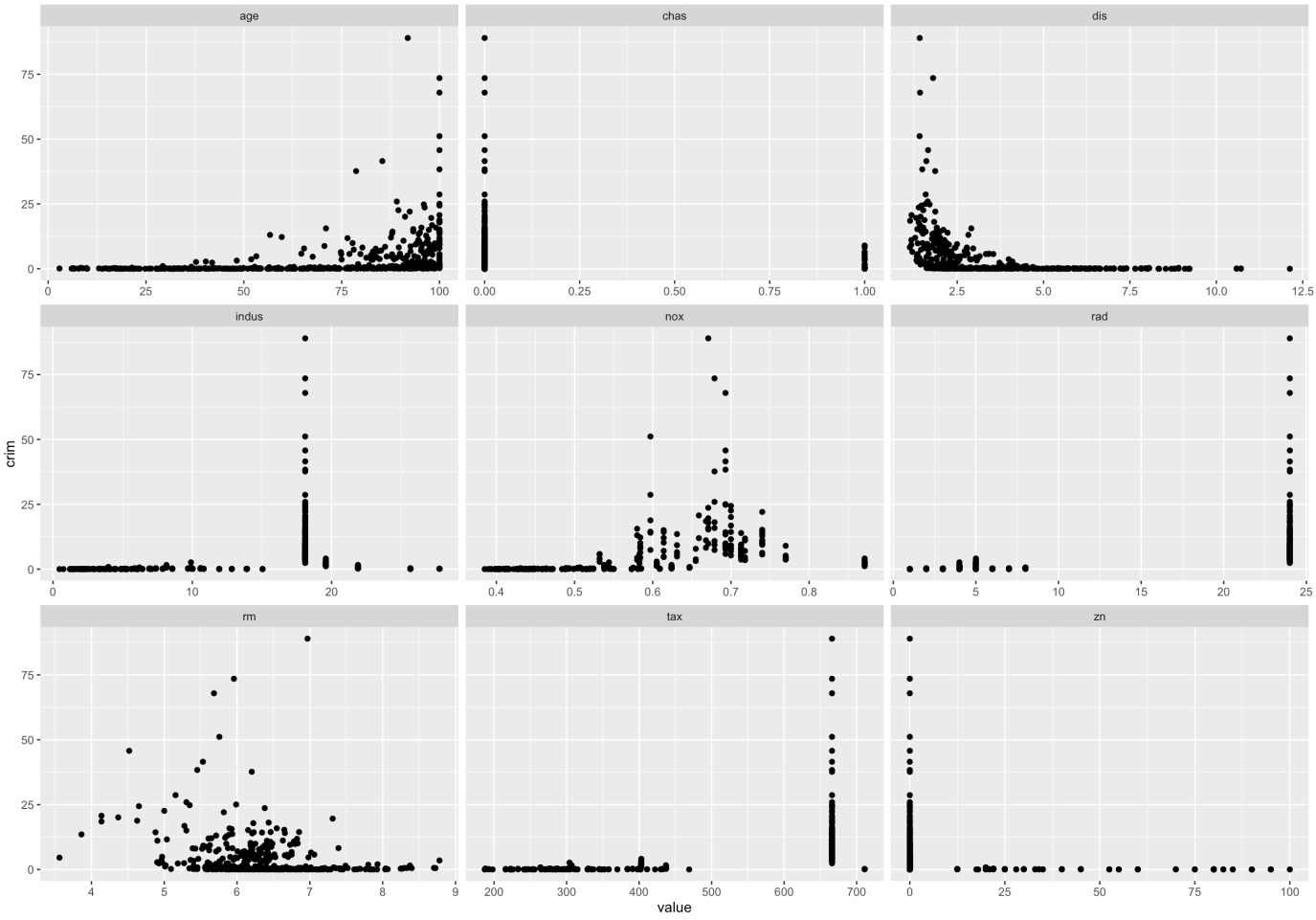


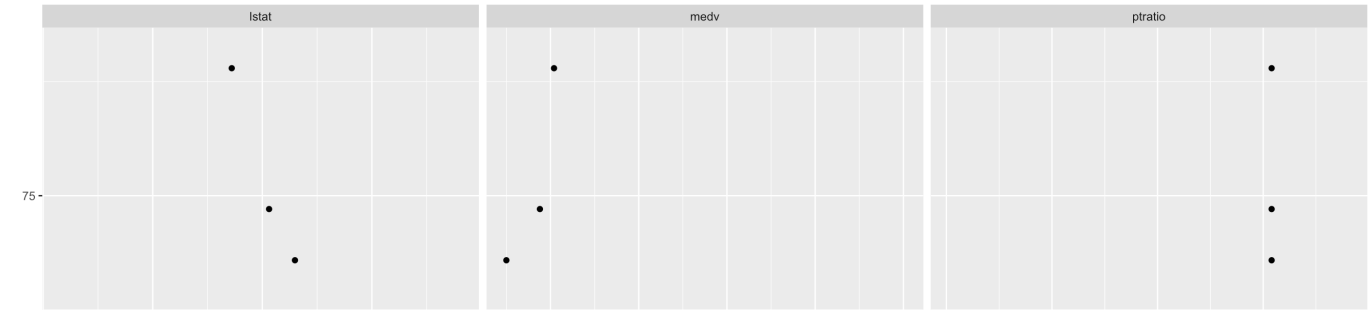
Bivariate Distribution

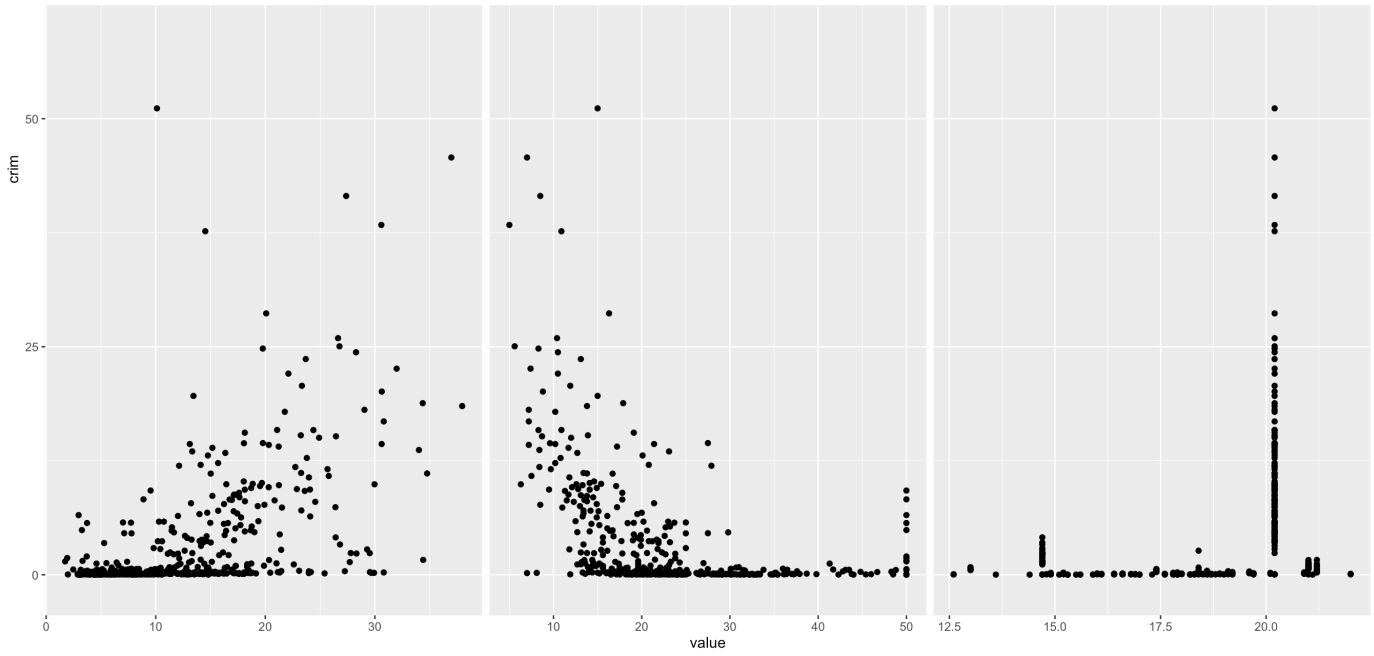
Boxplot (by crim)



Scatterplot (by crim)







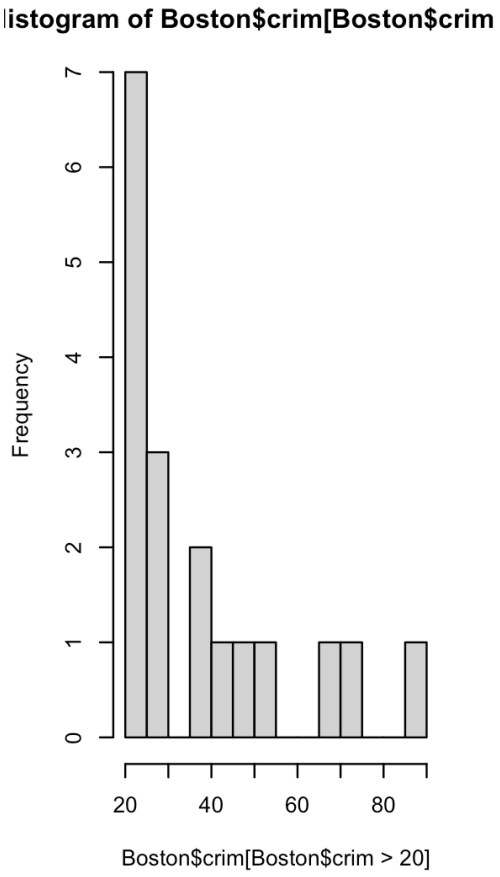
Page 2

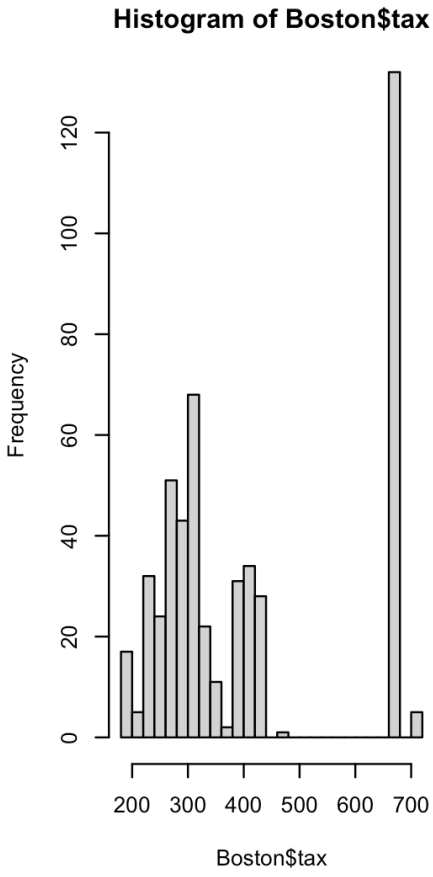
Hide

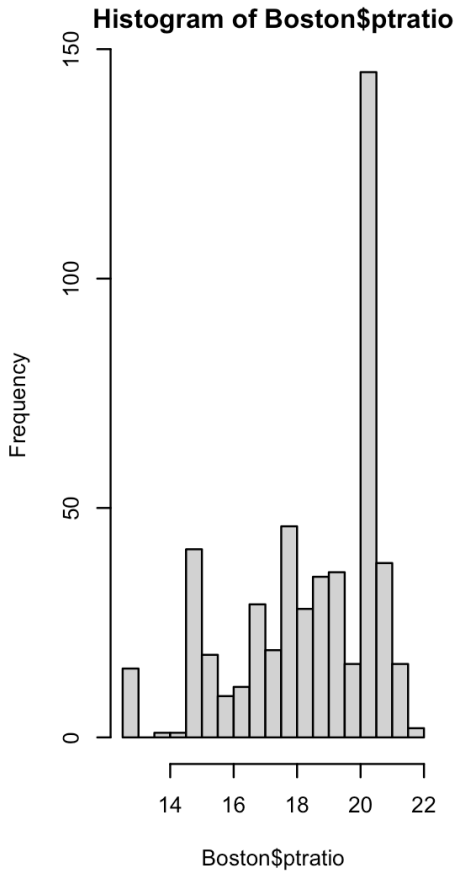
```
par(mfrow = c(1, 3))
hist(Boston$crim[Boston$crim > 20], breaks = 20)
hist(Boston$tax, breaks = 20)
```

Hide

```
hist(Boston$ptratio, breaks = 20)
```







Some places have much higher predictor values than other places. It's probably the wealthy areas, I haven't looked.

10e

> How many of the census tracts in this data set bound the Charles river?

Hide

```
nrow(Boston[Boston$chas == 1,])
```

```
[1] 35
```

10f

What is the median pupil-teacher ratio among the towns in this data set?

Hide

```
summary(Boston$ptratio)[3]
```

```
Median
 19.05
```

10g

Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

Hide

```
Boston[Boston$medv == min(Boston$medv),]
```

	crim <dbl>	zn <dbl>	indus <dbl>	chas <int>	nox <dbl>	rm <dbl>	age <dbl>	dis <dbl>	rad <int>
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24

2 rows | 1-10 of 13 columns

`crim` and `lstat` are the only variables that are very different from each other. I'm not sure which tracts these are so I can't comment too much. But I do like that the tax is 666.

10h

In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

Hide

```
# More than 7 rooms
nrow(Boston[Boston$rm > 7,])
```

```
[1] 64
```

Hide

```
# More than 8 rooms
nrow(Boston[Boston$rm > 8,])
```

```
[1] 13
```

Hide

```
summary(Boston[Boston$rm <= 8,])
```

```
      crim                zn              indus            chas              nox
rm
 Min.   : 0.00632   Min.   :  0.0   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850
Min.   :3.561
 1st Qu.: 0.08014   1st Qu.:  0.0   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490
1st Qu.:5.879
 Median : 0.24522   Median :  0.0   Median : 9.69   Median :0.00000   Median :0.5380
Median :6.185
 Mean   : 3.68986   Mean   : 11.3   Mean   :11.24   Mean   :0.06694   Mean   :0.5551
Mean   :6.230
 3rd Qu.: 3.77498   3rd Qu.: 12.5   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240
3rd Qu.:6.575
 Max.   :88.97620   Max.   :100.0   Max.   :27.74   Max.   :1.00000   Max.   :0.8710
Max.   :7.929
      age              dis              rad              tax            ptratio
lstat            medv
 Min.   :  2.9   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60   Mi
n.   : 1.73   Min.   : 5.00
 1st Qu.: 44.4   1st Qu.: 2.088   1st Qu.: 4.000   1st Qu.:280.0   1st Qu.:17.40   1s
t Qu.: 7.34   1st Qu.:16.70
 Median : 77.3   Median : 3.216   Median : 5.000   Median :334.0   Median :19.10   Me
dian :11.65   Median :21.00
 Mean   : 68.5   Mean   : 3.805   Mean   : 9.604   Mean   :410.4   Mean   :18.51   Me
an   :12.87   Mean   :21.96
 3rd Qu.: 94.3   3rd Qu.: 5.215   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3r
d Qu.:17.11   3rd Qu.:24.80
 Max.   :100.0   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00   Ma
x.   :37.97   Max.   :50.00
```

Hide

```
summary(Boston[Boston$rm > 8,])
```

```
      crim                zn              indus              chas              nox
rm
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000   Min.   :0.4161
Min.   :8.034
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000   1st Qu.:0.5040
1st Qu.:8.247
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000   Median :0.5070
Median :8.297
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538   Mean   :0.5392
Mean   :8.349
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000   3rd Qu.:0.6050
3rd Qu.:8.398
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000   Max.   :0.7180
Max.   :8.780
      age              dis              rad               tax           ptratio
lstat           medv
 Min.   : 8.40   Min.   :1.801   Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min
.   :2.47   Min.   :21.9
 1st Qu.:70.40   1st Qu.:2.288   1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st
Qu.:3.32   1st Qu.:41.7
 Median :78.30   Median :2.894   Median : 7.000   Median :307.0   Median :17.40   Med
ian :4.14   Median :48.3
 Mean   :71.54   Mean   :3.430   Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mea
n   :4.31   Mean   :44.2
 3rd Qu.:86.50   3rd Qu.:3.652   3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd
Qu.:5.12   3rd Qu.:50.0
 Max.   :93.90   Max.   :8.907   Max.   :24.000   Max.   :666.0   Max.   :20.20   Max
.   :7.44   Max.   :50.0
```

I'm just going to focus on the age bit because Boston is an historic city and area. I'm not surprised that larger homes are generally older than those that are not as large. This is probably due to old mansions which are abundant in the area. Also, median value is way higher for larger homes which is not at all surprising. I wish I knew what the tracts actually were so I could comment on that specific area rather than just the data.