Statistical Machine Learning

Descriptive Statistics

Day 38

How many pairs of shoes?

Step 1: Translation to math/statistics

Predict Y = number of pairs of shoes

For: students who take Math 361 at Oregon Tech

Measure quality of predictions by MAE

Available Predictors

- What is your height in inches?
- What is your favorite number?
- How much do you enjoy visiting Crater Lake, on a scale of 1 to 9?
- How do you typically commute to your classes at OIT? (Walk, car or public transit?)
- How many pairs of shoes do you own?
- Do you prefer the taste of coffee or coca cola? (1 = coffee, 2 = coca cola)
- Do you like milk chocolate better than dark chocolate? (Yes or no)
- What is your favorite color?

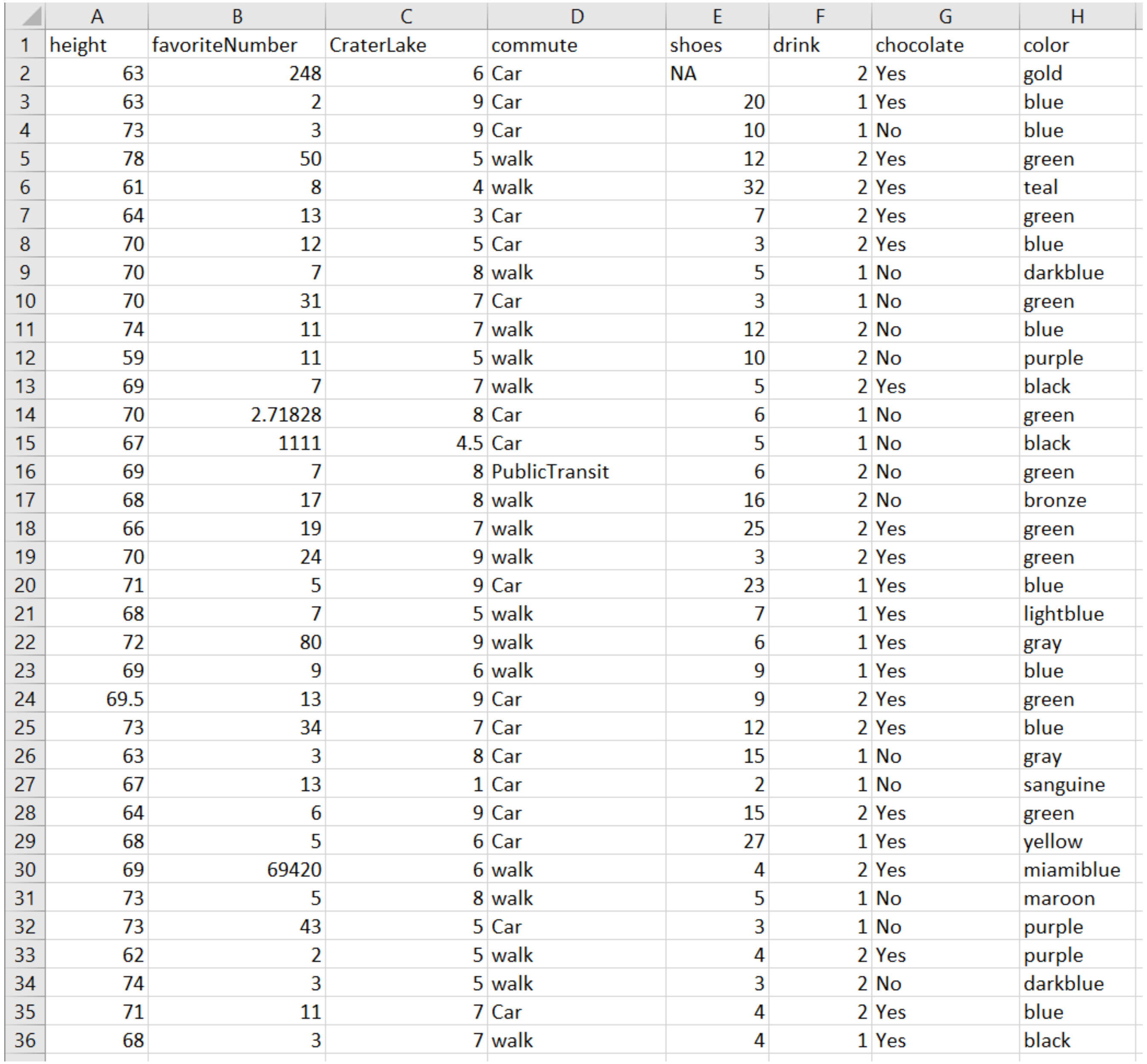
Step 2

Cleaned dataset:

All values are valid
Standard format per variable
    Colors have been standardized
“mistakes” corrected or set to NA

Transformed dataset

One variable per column
One example per row



	A	B	C	D	E	F	G	H
1	height	favoriteNumber	CraterLake	commute	shoes	drink	chocolate	color
2	63	248	6	Car	NA	2	Yes	gold
3	63	2	9	Car	20	1	Yes	blue
4	73	3	9	Car	10	1	No	blue
5	78	50	5	walk	12	2	Yes	green
6	61	8	4	walk	32	2	Yes	teal
7	64	13	3	Car	7	2	Yes	green
8	70	12	5	Car	3	2	Yes	blue
9	70	7	8	walk	5	1	No	darkblue
10	70	31	7	Car	3	1	No	green
11	74	11	7	walk	12	2	No	blue
12	59	11	5	walk	10	2	No	purple
13	69	7	7	walk	5	2	Yes	black
14	70	2.71828	8	Car	6	1	No	green
15	67	1111	4.5	Car	5	1	No	black
16	69	7	8	PublicTransit	6	2	No	green
17	68	17	8	walk	16	2	No	bronze
18	66	19	7	walk	25	2	Yes	green
19	70	24	9	walk	3	2	Yes	green
20	71	5	9	Car	23	1	Yes	blue
21	68	7	5	walk	7	1	Yes	lightblue
22	72	80	9	walk	6	1	Yes	gray
23	69	9	6	walk	9	1	Yes	blue
24	69.5	13	9	Car	9	2	Yes	green
25	73	34	7	Car	12	2	Yes	blue
26	63	3	8	Car	15	1	No	gray
27	67	13	1	Car	2	1	No	sanguine
28	64	6	9	Car	15	2	Yes	green
29	68	5	6	Car	27	1	Yes	yellow
30	69	69420	6	walk	4	2	Yes	miamiblue
31	73	5	8	walk	5	1	No	maroon
32	73	43	5	Car	3	1	No	purple
33	62	2	5	walk	4	2	Yes	purple
34	74	3	5	walk	3	2	No	darkblue
35	71	11	7	Car	4	2	Yes	blue
36	68	3	7	walk	4	1	Yes	black

Step 3: Understanding the dataset

Tools for describing a dataset:

- Graphs
- Numerical Summaries (a.k.a computing a “statistic”)

Numerical summaries

Univariate

Is your variable quantitative or qualitative?

measures of center: mean or median

measures of spread: min, max, IQR, SD

Proportion or counts of every possible value

Also, number of missing values, Number of unique values

Bivariate

2 quantitative variables – correlation measures strength of linear relationship, **if** the variables are linearly related.

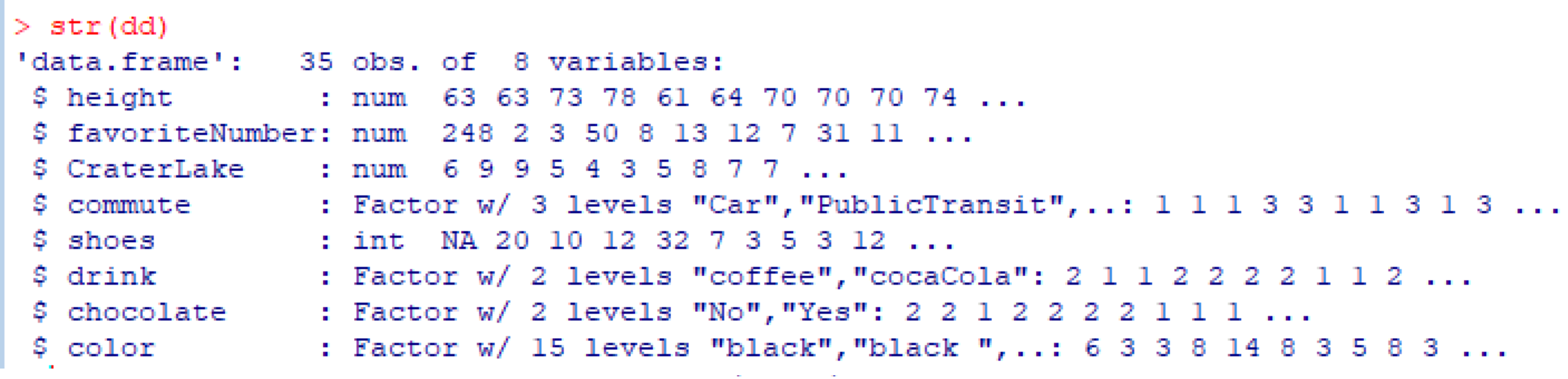
2 qualitative variables – conditional probabilities

1 quantitative, 1 qualitative – conditional measures of center, spread

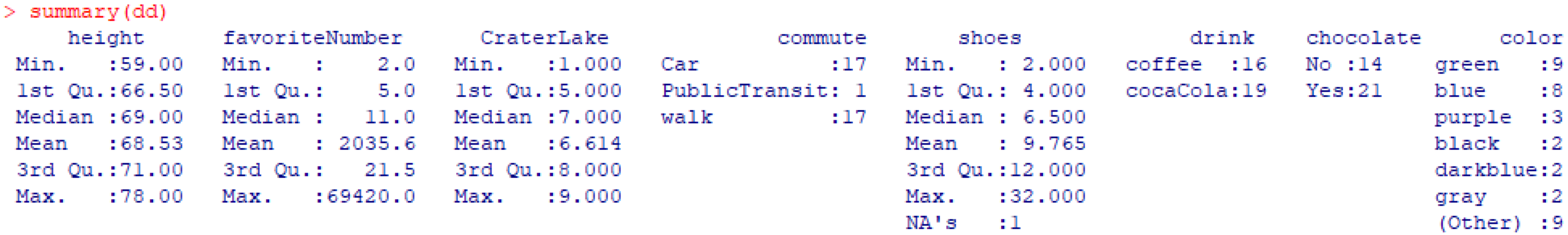
Univariate Summaries

Easy when data is read into a **data.frame** in R or a **pandas** data.frame in Python and any missing values are labeled “NA”

Check that the variables are saved as quantitative or qualitative as appropriate, then ask for **a summary** of the dataset.

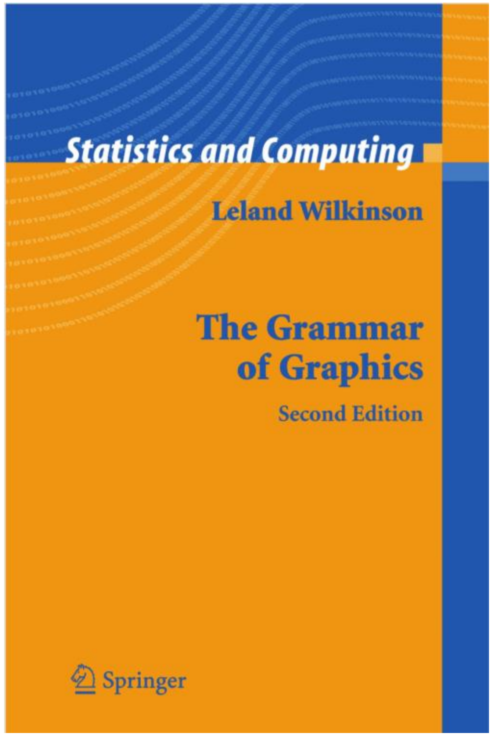


```
> str(dd)
'data.frame':   35 obs. of  8 variables:
 $ height        : num  63 63 73 78 61 64 70 70 70 74 ...
 $ favoriteNumber: num  248 2 3 50 8 13 12 7 31 11 ...
 $ CraterLake    : num  6 9 9 5 4 3 5 8 7 7 ...
 $ commute       : Factor w/ 3 levels "Car","PublicTransit",..: 1 1 1 3 3 1 1 3 1 3 ...
 $ shoes         : int  NA 20 10 12 32 7 3 5 3 12 ...
 $ drink         : Factor w/ 2 levels "coffee","cocaCola": 2 1 1 2 2 2 2 1 1 2 ...
 $ chocolate     : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 1 1 1 ...
 $ color         : Factor w/ 15 levels "black","black ",..: 6 3 3 8 14 8 3 5 8 3 ...
```



```
> summary(dd)
     height      favoriteNumber      CraterLake                 commute         shoes               drink     chocolate        color   
 Min.   :59.00   Min.   :    2.0   Min.   :1.000   Car           :17   Min.   : 2.000   coffee  :16   No :14    green   :9  
 1st Qu.:66.50   1st Qu.:    5.0   1st Qu.:5.000   PublicTransit: 1   1st Qu.: 4.000   cocaCola:19   Yes:21    blue    :8  
 Median :69.00   Median :   11.0   Median :7.000   walk          :17   Median : 6.500                            purple  :3  
 Mean   :68.53   Mean   : 2035.6   Mean   :6.614                       Mean   : 9.765                            black   :2  
 3rd Qu.:71.00   3rd Qu.:   21.5   3rd Qu.:8.000                       3rd Qu.:12.000                            darkblue:2  
 Max.   :78.00   Max.   :69420.0   Max.   :9.000                       Max.   :32.000                            gray    :2  
                                                                       NA's   :1                                 (Other) :9  
```

A Layered Grammar of Graphics



Statistics and Computing

Leland Wilkinson

The Grammar of Graphics

Second Edition

Springer

Use R!

Hadley Wickham

ggplot2

Elegant Graphics for Data Analysis

Springer

A language for describing the key features of statistical graphs:

“In brief, the grammar tells us that a statistical graphic is a mapping from data to aesthetic attributes (colour, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system. Faceting can be used to generate the same plot for different subsets of the dataset. It is the combination of these independent components that make up a graphic”

-Hadley Wickham, ggplot2

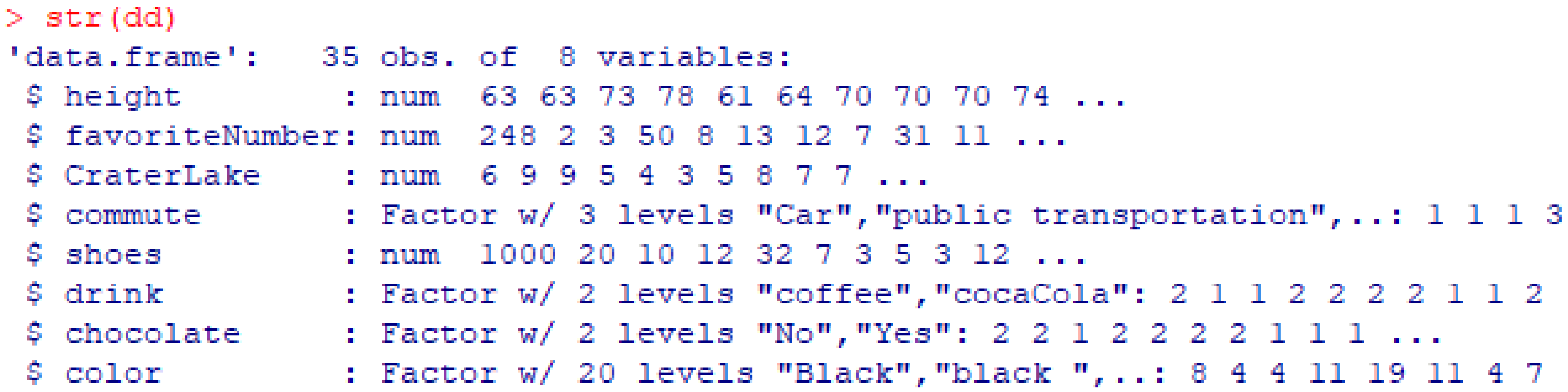
Implemented for R in the **ggplot2** library for data.frame objects and
For python in the **plotnine** library for pandas objects

A few ideas I’ve found useful:

Build a graph up in “layers”

- Map variables in the data to “aesthetic” attributes
- Choice of geometric objects to draw
- Apply statistical transformations to summarize the dataset (optional)
- Faceting to create the same type of graph for different part of the dataset (optional)

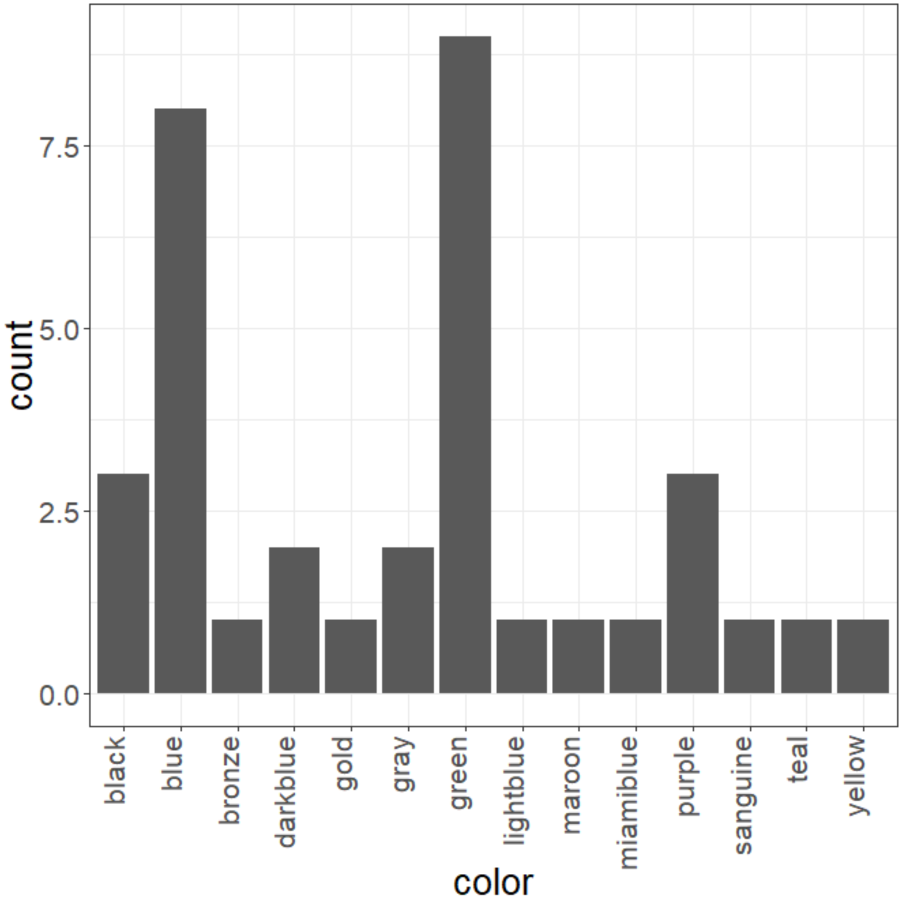
Math 361 Surveys



```
> str(dd)
'data.frame':   35 obs. of  8 variables:
 $ height        : num  63 63 73 78 61 64 70 70 70 74 ...
 $ favoriteNumber: num  248 2 3 50 8 13 12 7 31 11 ...
 $ CraterLake    : num  6 9 9 5 4 3 5 8 7 7 ...
 $ commute       : Factor w/ 3 levels "Car","public transportation",..: 1 1 1 3
 $ shoes         : num  1000 20 10 12 32 7 3 5 3 12 ...
 $ drink         : Factor w/ 2 levels "coffee","cocaCola": 2 1 1 2 2 2 2 1 1 2
 $ chocolate     : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 1 1 1 ...
 $ color         : Factor w/ 20 levels "Black","black ",..: 8 4 4 11 19 11 4 7
```

A barchart of favorite colors

```
ggplot(dd, aes(x=color))+
    geom_bar()
```

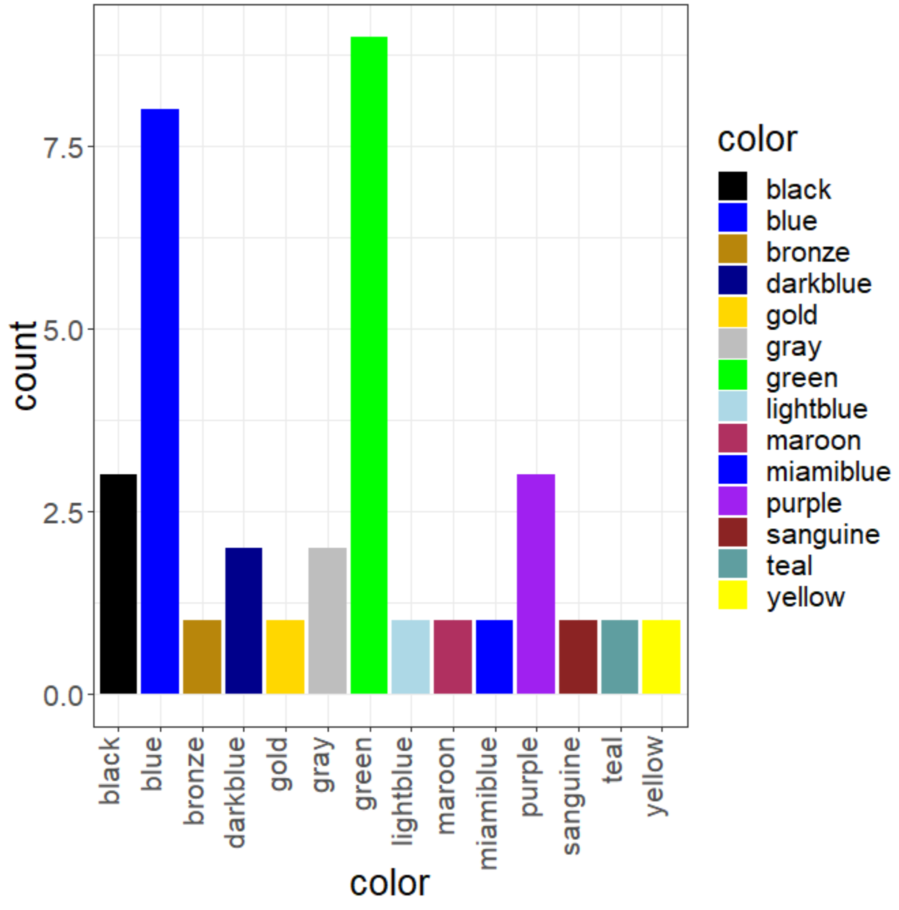


A **barchart** can be used to visualize a single qualitative variable

- Map the variable values to the x-axis
- Geom = bar
- Statistical Transform = count

A barchart of favorite colors

```
ggplot(dd, aes(x=color, fill=color))+
    geom_bar()
```

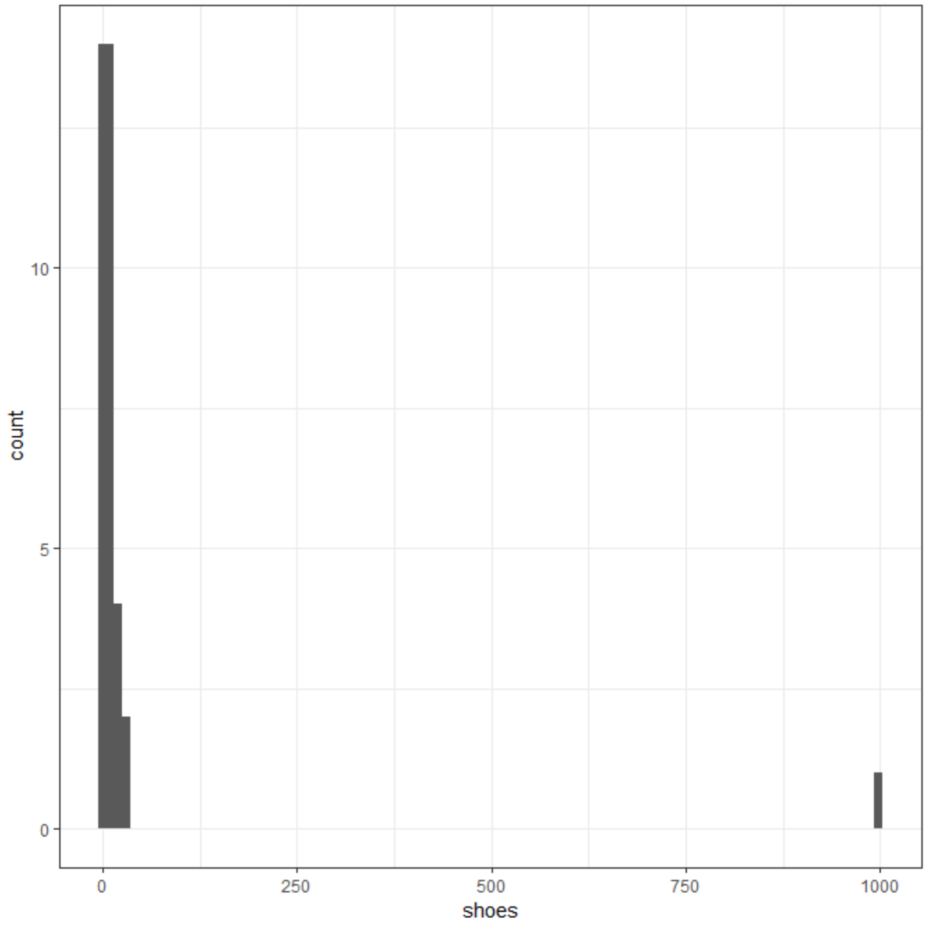


A **barchart** can be used to visualize a single qualitative variable

- Map the variable values to the x-axis (and fill color)
- Geom = bar
- Statistical Transform = count

Histogram of Number of pairs of shoes

```
ggplot(dd, aes(x=shoes))+geom_histogram(bins=100)+theme_bw()
```

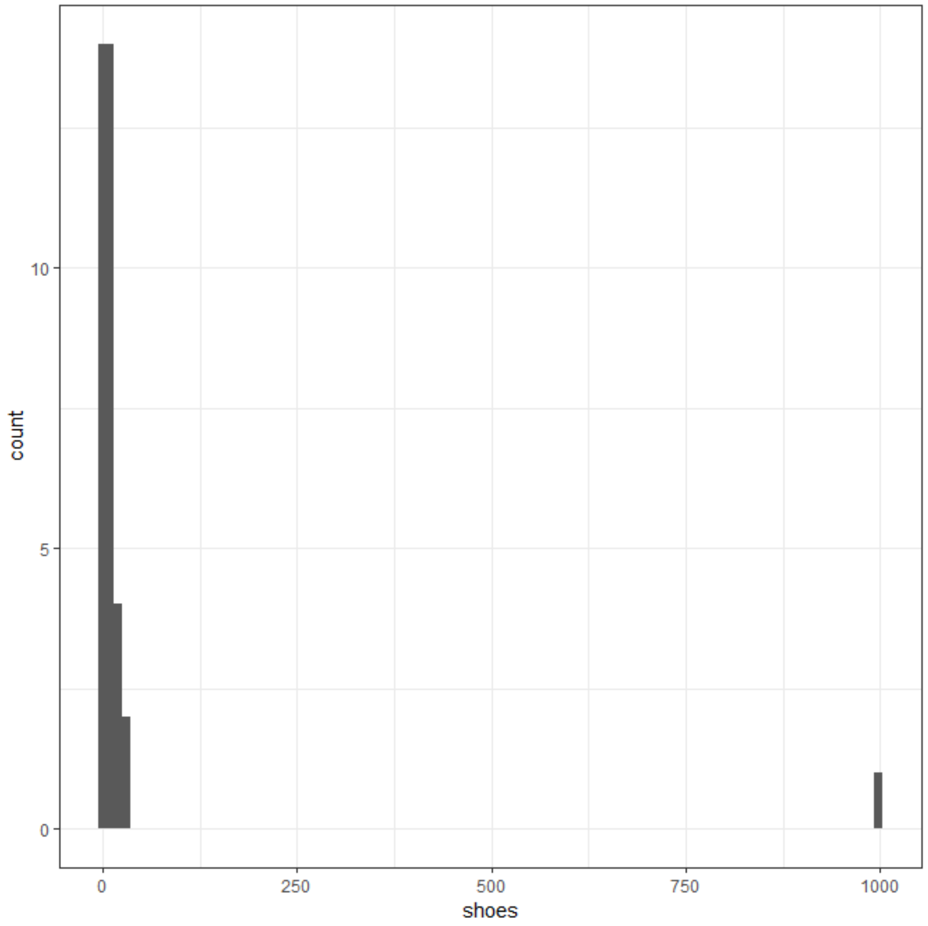


A **histogram** can be used to visualize a single quantitative variable

- Map the variable values to the x-axis
- Geom = bar
- Statistical Transform = bin

Histogram of Number of pairs of shoes

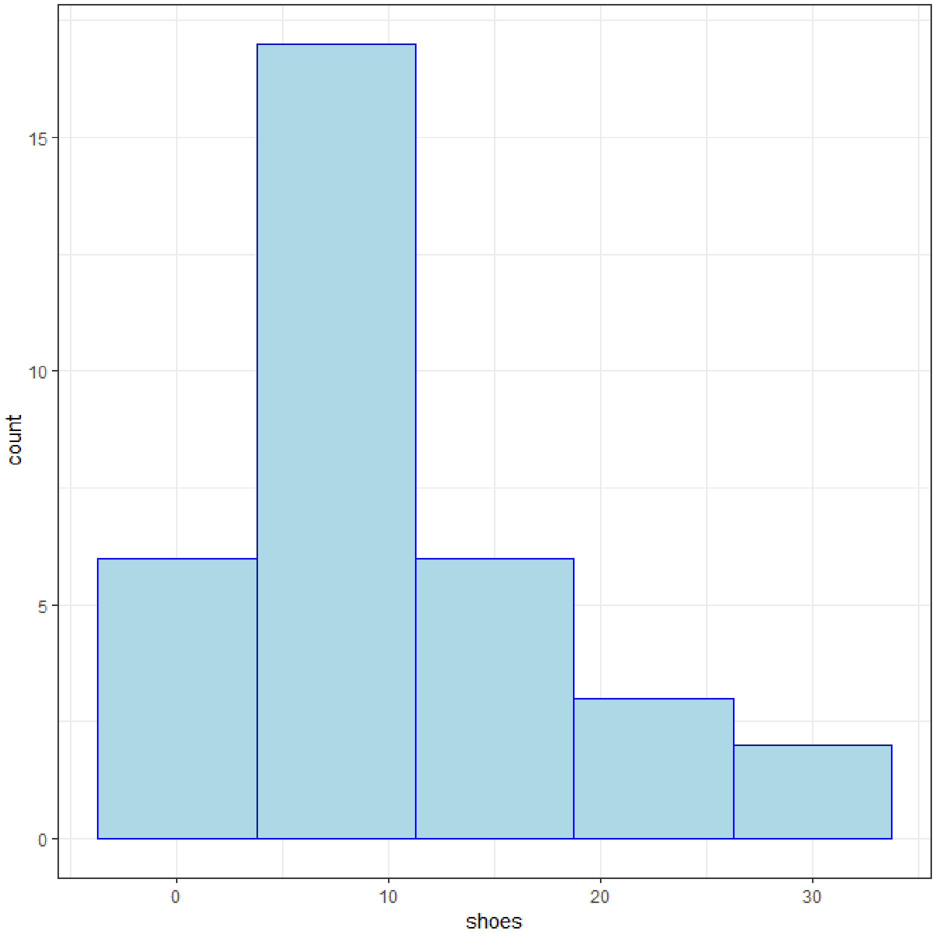
```
ggplot(dd, aes(x=shoes))+geom_histogram(bins=100)+theme_bw()
```



A **histogram** can be used to visualize a single quantitative variable

- Map “shoes” to the x-axis
- Geom = bar
- Statistical Transform = bin

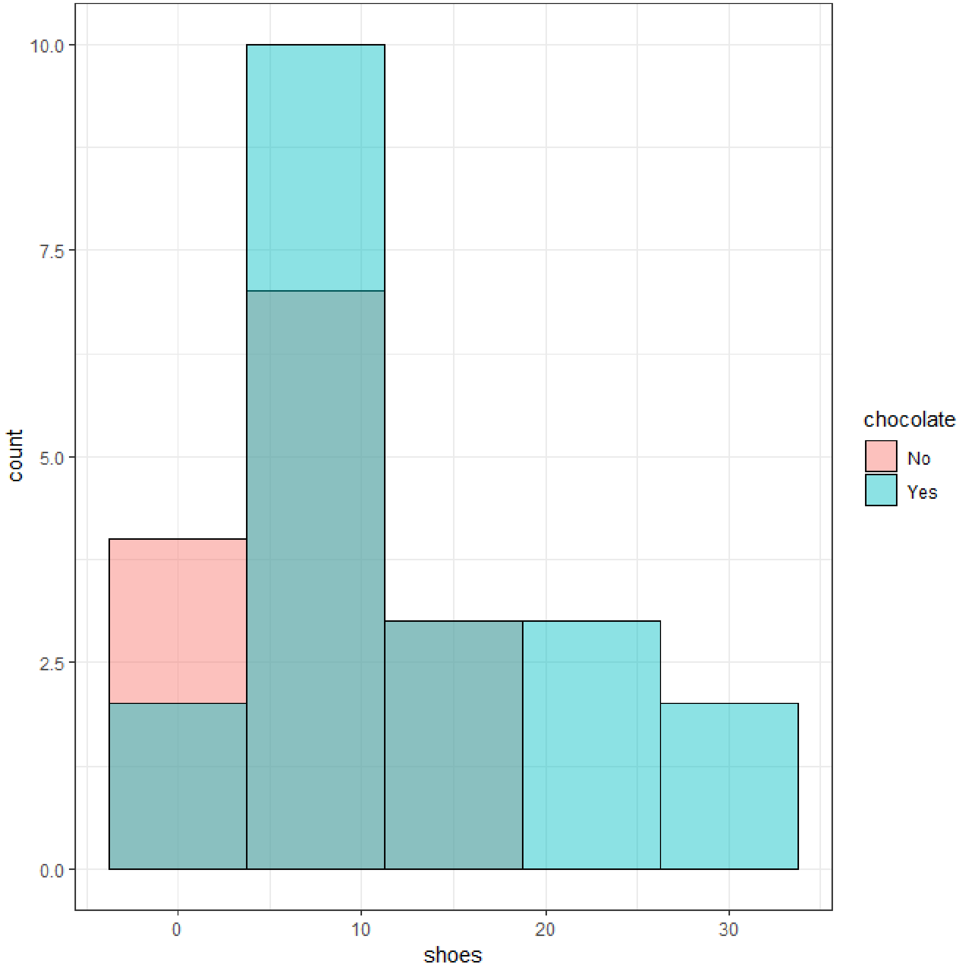
Histogram of number of pairs of shoes, with one outlier (1000 pairs) removed



```
ggplot(dd[dd$shoes<1000,], aes(x=shoes))+
       geom_histogram(bins=5,color="blue", fill="lightblue")+
       theme_bw()
```

Histograms of Shoes by Chocolate Preference

```
ggplot(dd[dd$shoes<1000,], aes(x=shoes, fill=chocolate))+
       geom_histogram(bins=5, color="black", alpha=0.45, position="identity")+
       theme_bw()
```

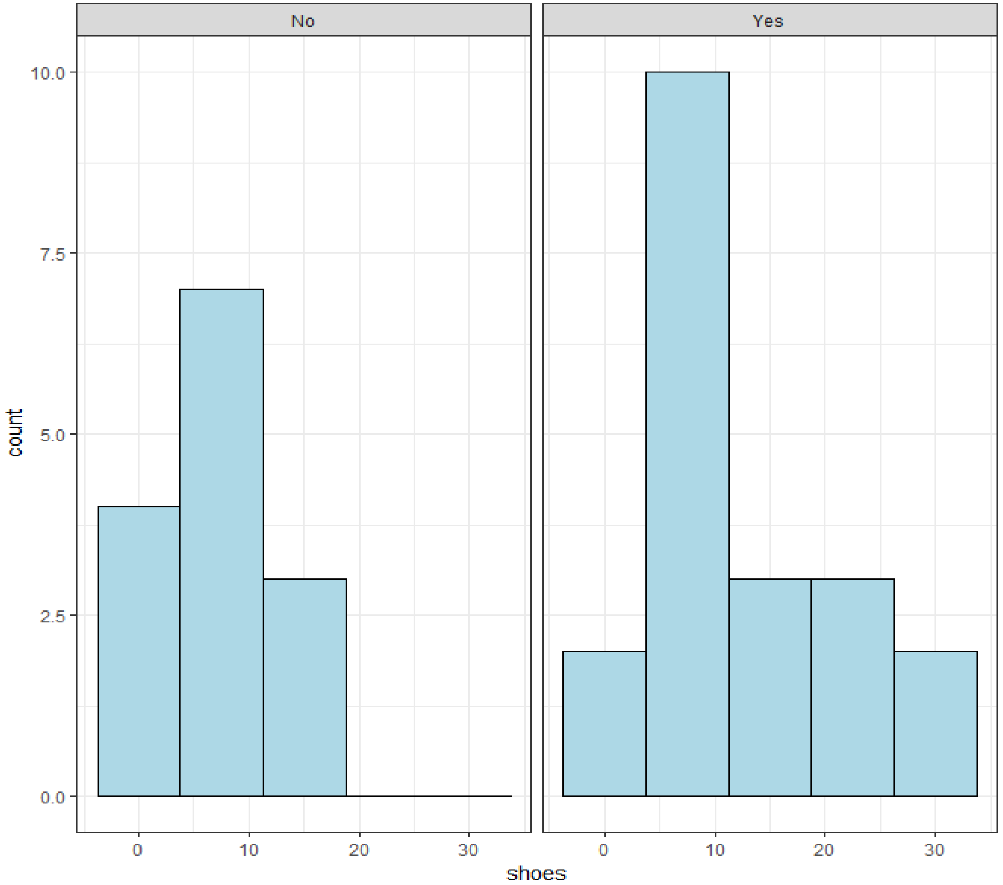


Two histograms can be used to visualize a qualitative and a quantitative variable

- Map “shoes” to the x-axis
- Map “chocolate” to color
- Geom = bar
- Statistical Transform = bin

Histograms of Shoes by Chocolate Preference

```
ggplot(dd[dd$shoes<1000,], aes(x=shoes))+
       geom_histogram(bins=5, fill="lightblue", color="black")+
       facet_wrap(~chocolate)+
       theme_bw()
```

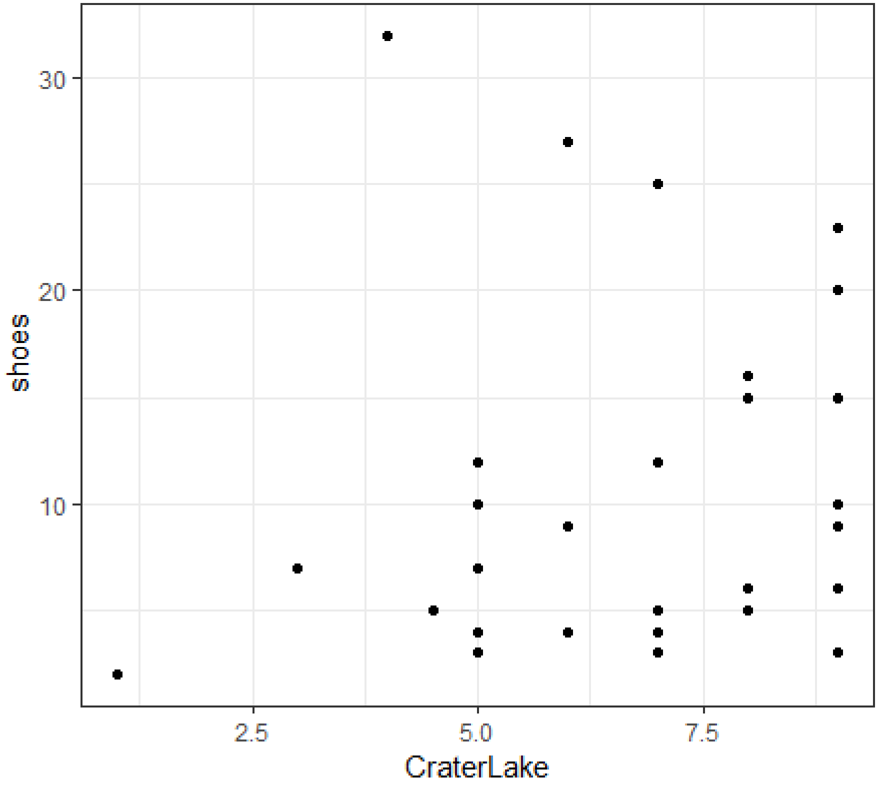


Two histograms can be used to visualize a qualitative and a quantitative variable:

- Map “shoes” to the x-axis
- Geom = bar
- Statistical Transform = bin
- Facet = by “chocolate”

Scatterplot of Crater Lake Rating and number of pairs of shoes

```
ggplot(dd[dd$shoes<1000,], aes(x=CraterLake, y=shoes))+
       geom_point()+
       theme_bw()
```

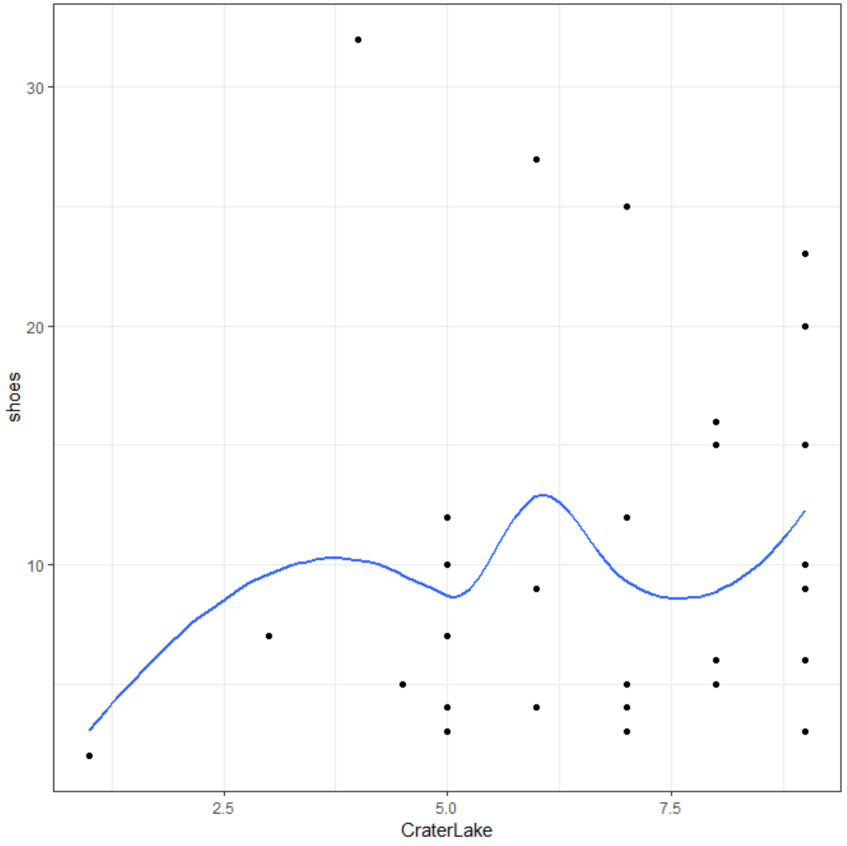


Scatterplot for two quantitative variables:

- Map “craterlake” to x-axis
- Map “shoes” to y-axis
- Geom – point
- Stat – identity

Scatterplot of Crater Lake Rating and number of pairs of shoes

```
ggplot(dd[dd$shoes<1000,], aes(x=CraterLake, y=shoes))+
       geom_point()+
       stat_smooth(se=FALSE)+
       theme_bw()
```

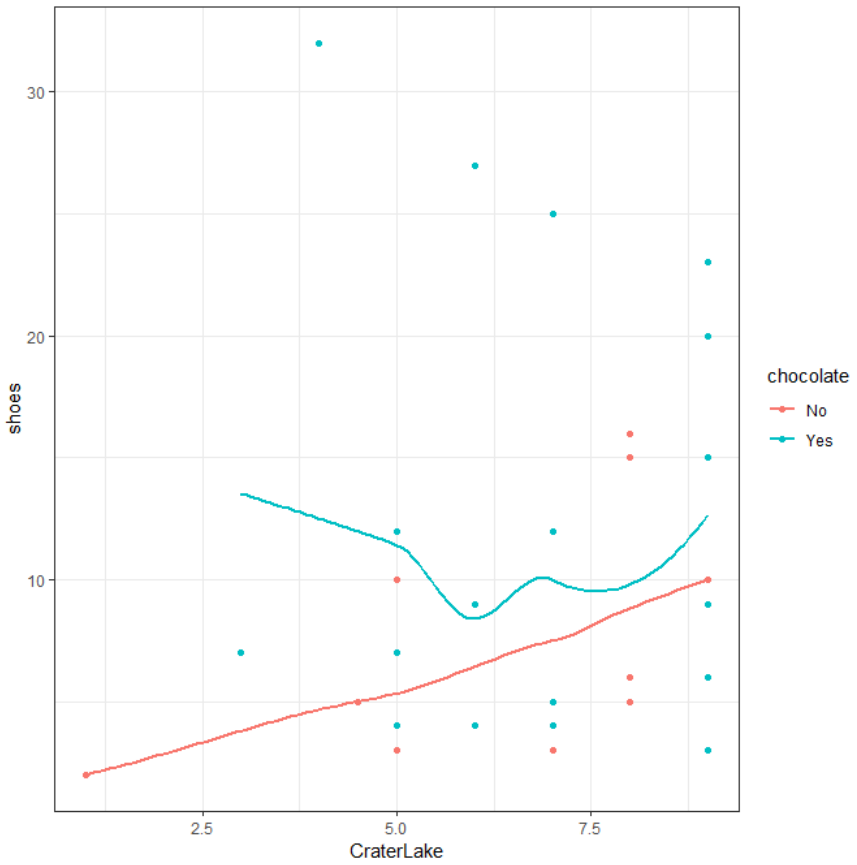


Scatterplot for two quantitative variables:

- Map “craterlake” to x-axis
- Map “shoes” to y-axis
- Geom – point
- Stat – smoothed fit

Scatterplot of Crater Lake Rating and number of pairs of shoes by chocolate preference

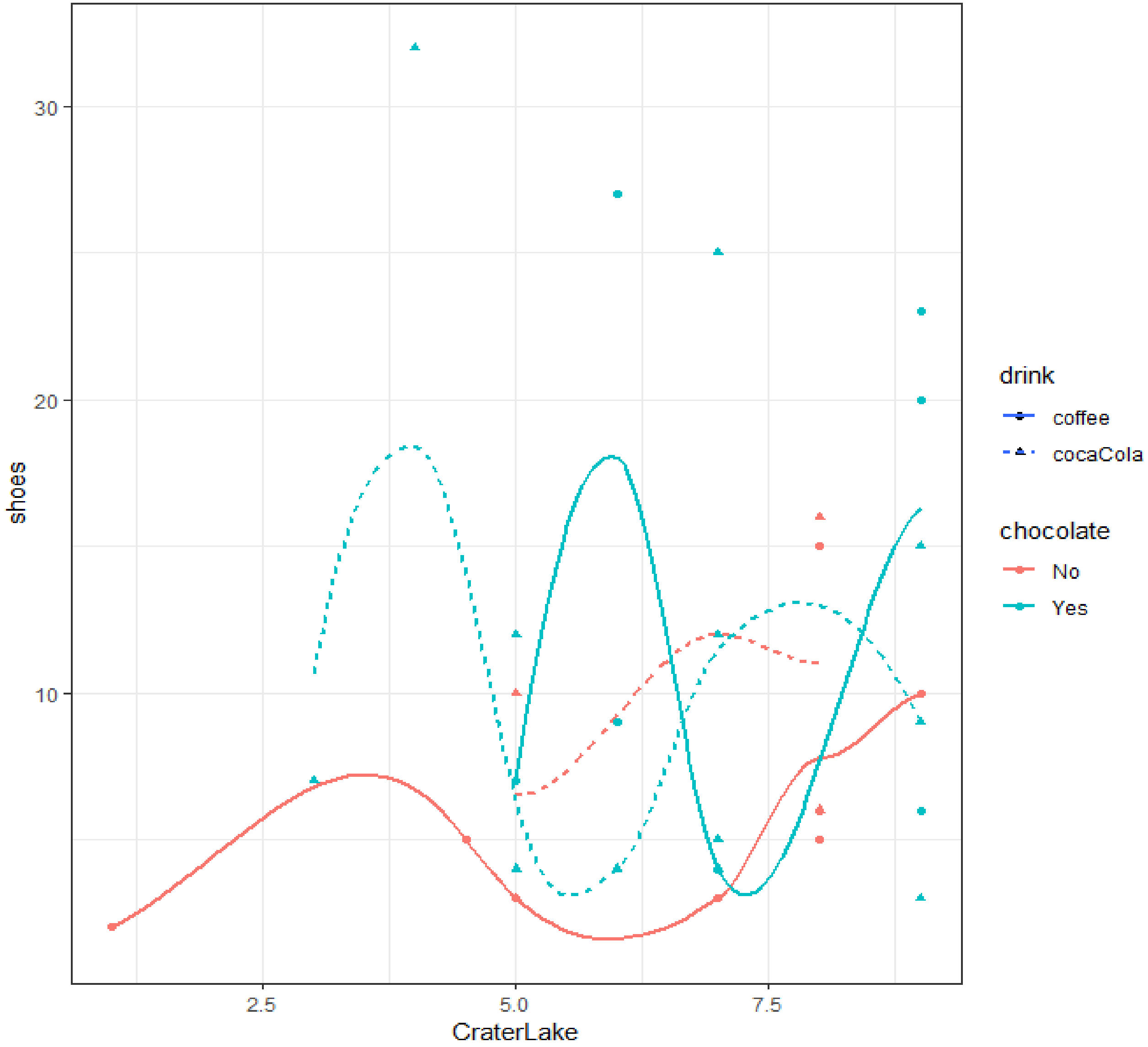
```
ggplot(dd[dd$shoes<1000,], aes(x=CraterLake, y=shoes, color=chocolate))+
       geom_point()+
       stat_smooth(se=FALSE)+
       theme_bw()
```



Scatterplot for two quantitative variables and one qualitative variable:

- Map “craterlake” to x-axis
- Map “shoes” to y-axis
- Map “chocolate” to color
- Geom – point
- Stat – smoothed fit

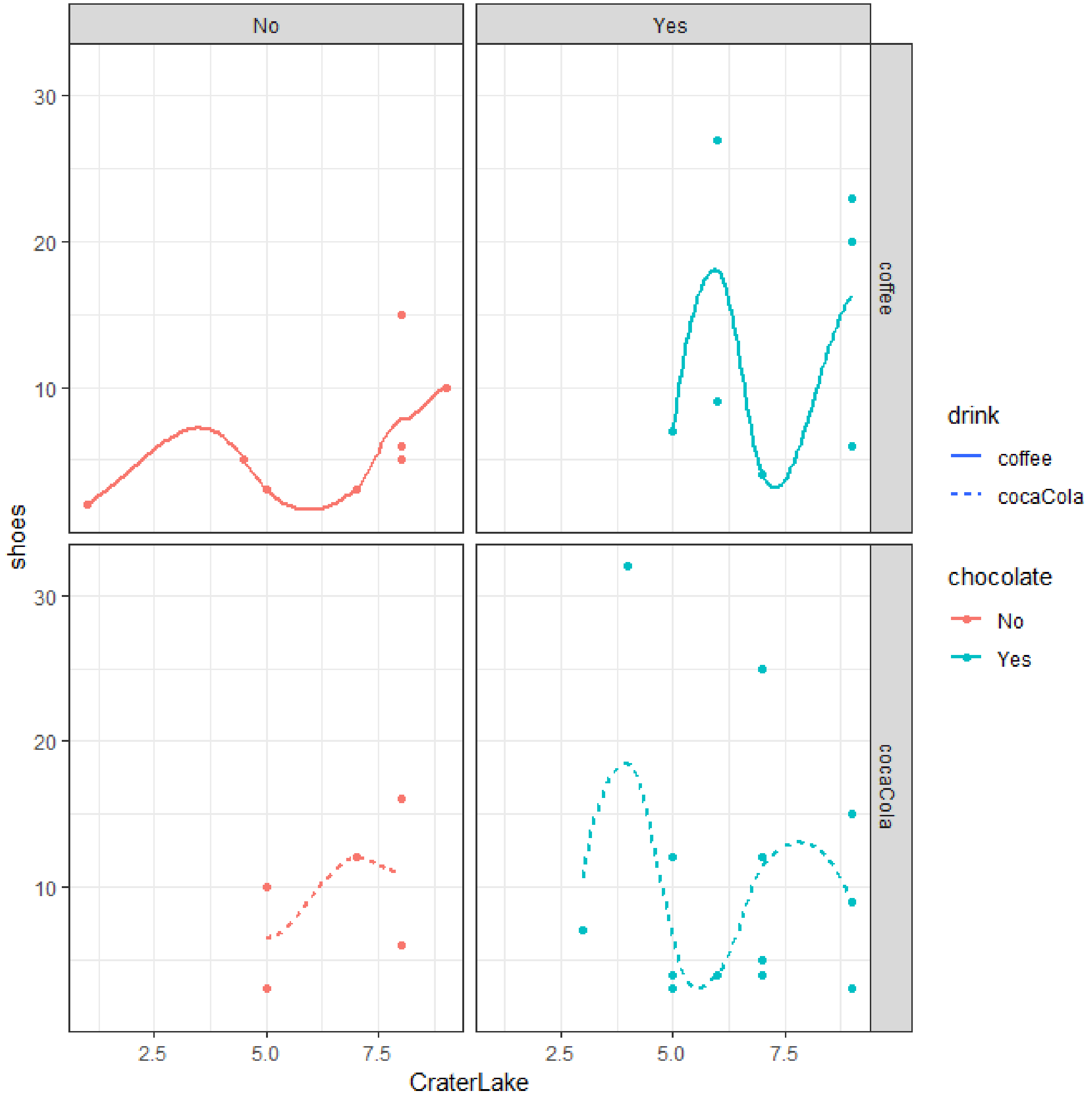
Scatterplot of Crater Lake Rating and number of pairs of shoes by chocolate preference and drink preference….



Scatterplot for two quantitative variables and two qualitative variables

- Map “craterlake” to x-axis
- Map “shoes” to y-axis
- Map “chocolate” to color
- Map “drink” to linetype, point shape
- Geom – point
- Stat – smoothed fit

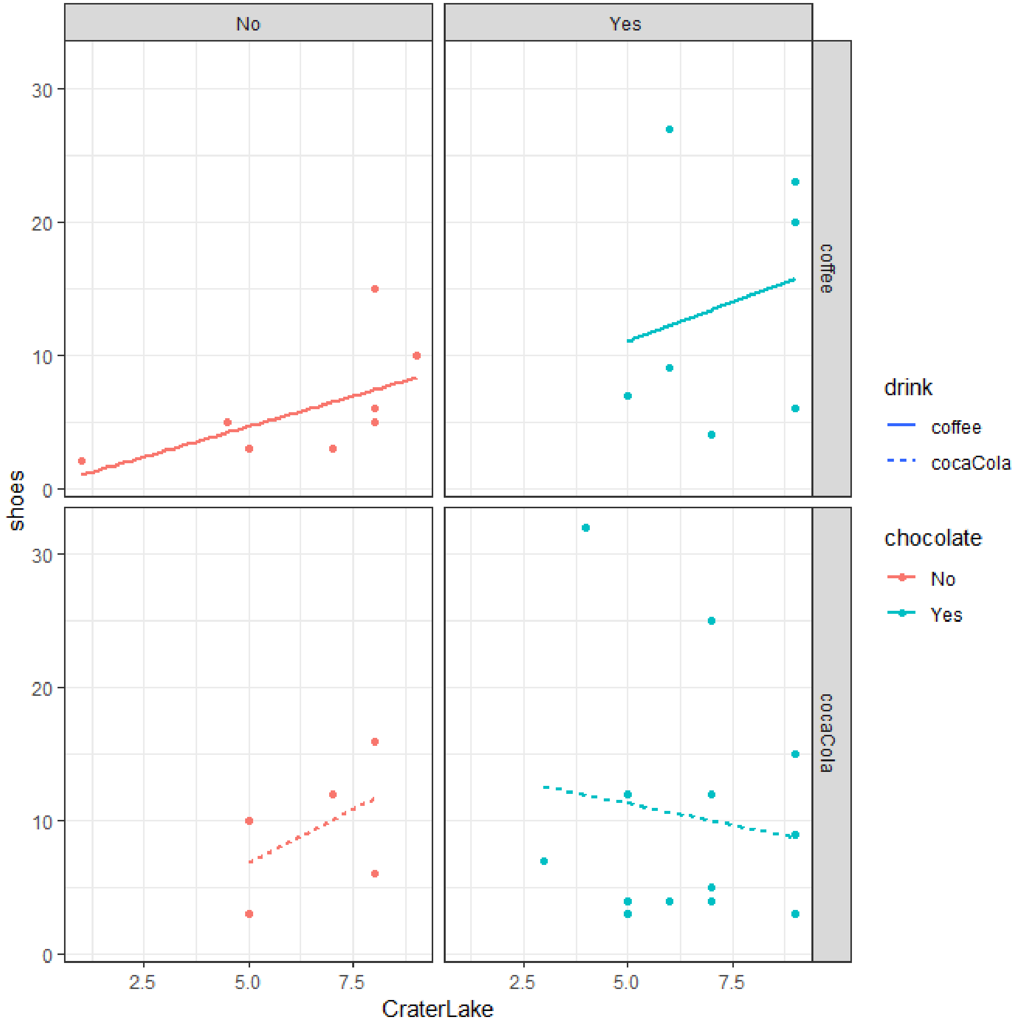
Scatterplot of Crater Lake Rating and number of pairs of shoes by chocolate preference and drink preference….



Scatterplot for two quantitative variables and two qualitative variables

- Map “craterlake” to x-axis
- Map “shoes” to y-axis
- Map “chocolate” to color
- Map “drink” to linetype
- Geom – point
- Stat – smoothed fit
- Facet by drink and chocolate

Scatterplot of Crater Lake Rating and number of pairs of shoes by chocolate preference and drink preference….



Scatterplot for two quantitative variables and two qualitative variables

- Map "craterlake" to x-axis
- Map "shoes" to y-axis
- Map "chocolate" to color
- Map "drink" to linetype
- Geom – point
- Stat – linear regression fit
- Facet by drink and chocolate

Possible aesthetics for the point geom

A variable can be mapped to:

- Distance along X-axis
- Distance along Y-axis
- Alpha (transparency)
- Color/fill
- Group
- Shape
- Size
- Stroke

https://ggplot2.tidyverse.org/articles/ggplot2-specs.html

Making graphs is fun…but what’s our goal?

- Do any of the possible predictors seem to have a relationship with Y = number of pairs of shoes? If so, what is the form of the relationship?

For each possible predictor X, make a graph with Y

If Y is quantitative, use geom = (point) and map Y to the y-axis and either

map a quantitative X to the x-axis, with stat=(smooth or linear fit)

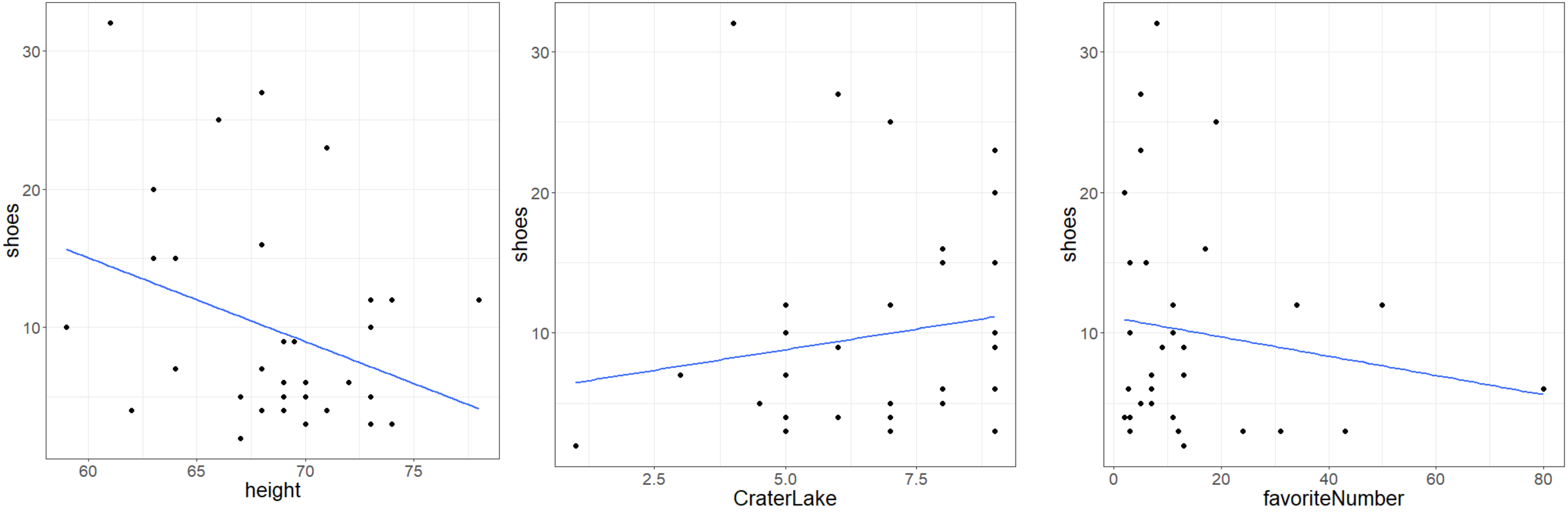
or

map a qualitative X to jittered x-axis AND color, with stat= (5-number summary, a.k.a “boxplot”)

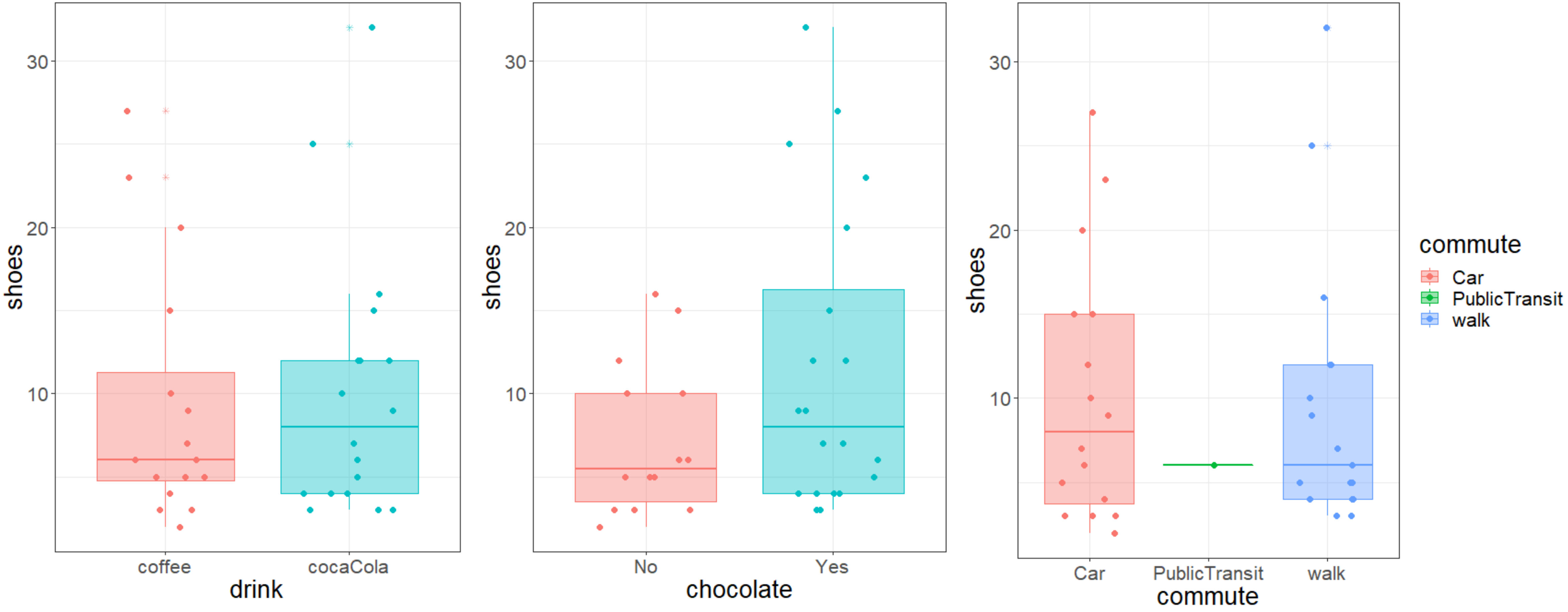
- Does a pair of possible predictors seem to have a relationship with Y?

Add another layer to the above plot with the second possible predictor mapped to shape, size, color and/or facet

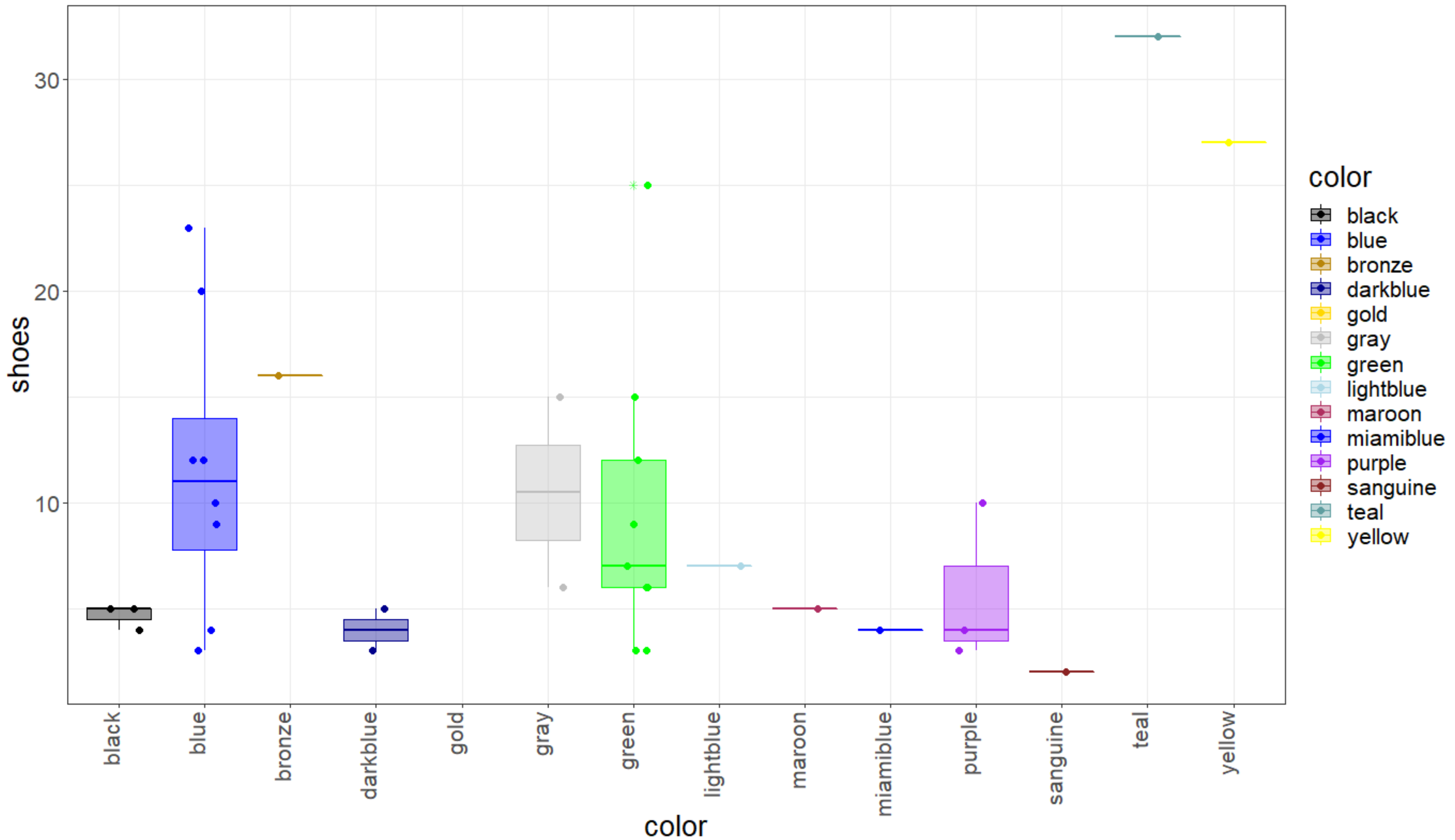
Shoes vs height, crater lake and favorite number



Shoes vs. drink, chocolate, commute



Shoes vs. favorite color



What did I learn about the individual predictors and their relationship with Y = # pairs of shoes?

- It looks like drink preference (cola or coffee) is the most related to number of pairs of shoes

- Not enough students to see a relationship with colors – could try combining colors into larger categories, i.e. warm vs. cool.

- very weak relationships with the other possible predictors

With so little data, maybe linear regression, with lasso or ridge will work best

- No transformations or penalized spline needed.

- Maybe interactions – look for them manually or try a neural net.

Does a person have heart disease?

Step 1: Transform question to math/statistics

Predict Y = 1 for heart disease, 0 if no heart disease

Who should the model work for? Americans in 1980 who visited the Cleveland clinic

Desired Quality of Predictions: False Positive Rate < 20%

True Positive Rate > 98%

13 **possible predictors** available in the dataset:

https://www.kaggle.com/ronitf/heart-disease-uci

Possible Predictors

age: age in years
sex: sex (1 = male; 0 = female)

cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic

trestbps: resting blood pressure (in mm Hg on admission to the hospital)
chol: serum cholestoral in mg/dl

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach: maximum heart rate achieved
exang: exercise induced angina (1 = yes; 0 = no)
oldpeak = ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping

ca: number of major vessels (0-3) colored by flourosopy
thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Y is diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing

Use the grammar of graphics to decide which graphs to create

We’re interested in the relationships of a binary Y with quantitative or qualitative Xs. For both situations:

- Choose a geom (bar, point, line…)
- Map variables to aesthetics (i.e. x or y axes, color, shape, size…)
- Add statistical transform (optional, i.e. counts, proportions, linear fit…)
- Facet (the same type of graph for different parts of the dataset) (optional)