THE UNIVERSITY OF WESTERN AUSTRALIA

School of Computer Science and Software Engineering

CITS4009 Introduction to Data Science

SEMESTER 2, 2017: CHAPTER 3 EXPLORING DATA

Chapter Objectives

- Using summary statistics to explore data
- Exploring data using visualization
- Finding problems and issues during data exploration

Organizing Data for Analysis

- In a database data is usually stored in normalized form to reduce redundancy: information about single user is spread across many small tables.
- This way makes adding easy but not ideal for analysis.
- You can join all the needed data into single table in database using SQL.

Using summary statistics to spot problems

The **`summary()`** command reports a variety of summary statistics on the numerical columns of the data frame, and count statistics on any categorical columns.

- Mean
- variance,
- median,
- min,
- max,
- quantile

```
> summary(custdata)
custid          sex
Min.   :   2068   F:440
1st Qu.: 345667   M:560
Median : 693403
Mean   : 698500
3rd Qu.:1044606
Max.   :1414286

is.employed         income
Mode :logical   Min.   : -8700
FALSE:73        1st Qu.: 14600
TRUE :599       Median : 35000
NA's :328       Mean   : 53505
                3rd Qu.: 67000
                Max.   :615000

marital.stat
Divorced/Separated:155
Married           :516
Never Married     :233
Widowed           : 96
```

The variable is.employed is missing for about a third of the data. The variable income has negative values, which are potentially invalid.

The summary of the data helps you quickly spot potential problems, like missing data or unlikely values

```
health.ins
Mode :logical
FALSE:159
TRUE :841
NA's :0

housing.type
Homeowner free and clear    :157
Homeowner with mortgage/loan:412
Occupied with no rent       : 11
Rented                      :364
NA's                        : 56

recent.move      num.vehicles
Mode :logical    Min.   :0.000
FALSE:820        1st Qu.:1.000
TRUE :124        Median :2.000
NA's :56         Mean   :1.916
                 3rd Qu.:2.000
                 Max.   :6.000
                 NA's   :56

age              state.of.res
Min.   :  0.0    California  :100
1st Qu.: 38.0    New York    : 71
Median : 50.0    Pennsylvania: 70
Mean   : 51.7    Texas       : 56
3rd Qu.: 64.0    Michigan    : 52
Max.   :146.7    Ohio        : 51
                 (Other)     :600
```

About 84% of the customers have health insurance.

The variables housing.type, recent.move, and num.vehicles are each missing 56 values.

The average value of the variable age seems plausible, but the minimum and maximum values seem unlikely. The variable state.of.res is a categorical variable; summary() reports how many customers are in each state (for the first few states).

Typical problems revealed by data summaries

At this stage you are looking for:

1. missing values
2. invalid values and outliers,
3. data ranges

1- Missing Values

- many modeling algorithms will, by default, quietly drop rows with missing values.
- What is the appropriate action with missing values?
  - Drop the data rows where you're missing.
  - Treat the `NAs` as a third employment category
  - Convert the bad data to a useful value.

2- Invalid values and outliers

- **Check data if it make sense or not.**
- **Invalid data** are simply bad data input
- **Outliers** are data points that fall well out of the range of where you expect the data to l

```
> summary(custdata$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  -8700   14600   35000   53500   67000
    Max.
 615000
```

Negative values for income could indicate bad data. They might also have a special meaning, like "amount of debt."

Either way, you should check how prevalent the issue is, and decide what to do: Do you drop the data with negative income? Do you convert negative values to zero?

```
> summary(custdata$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.
    0.0    38.0    50.0    51.7    64.0
    Max.
  146.7
```

Customers of age zero, or customers of an age greater than about 110 are outliers. They fall out of the range of expected customer values.

Outliers could be data input errors. They could be special sentinel values: zero might mean "age unknown" or "refuse to state." And some of your customers might be especially long-lived.

3- Data range

- How much the values in the data vary
- Is the data range wide? Is it narrow?

```
> summary(custdata$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  -8700   14600   35000   53500   67000
   Max.
 615000
```

Income ranges from zero to over half a million dollars; a very wide range.

- How narrow is too narrow" a data range?

Rely on information about the problem domain to judge if the data range is narrow, but a rough rule of thumb is the ratio of the standard deviation to the mean. If that ratio is very small, then the data isn't varying much.

- Factor that determines apparent data range is the ***unit*** of measurement.
- Checking units can prevent inaccurate results later.

3.2. Spotting problems using graphics and visualization

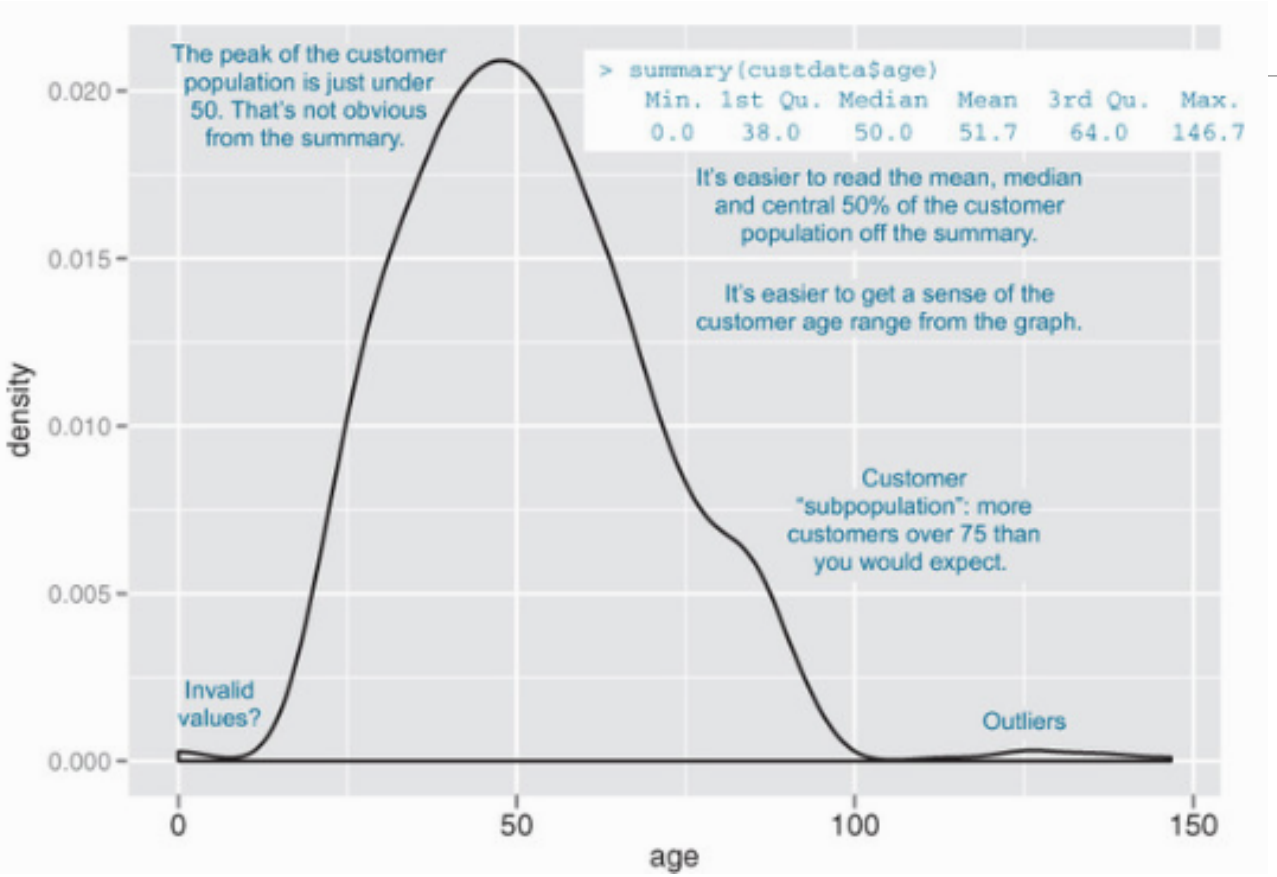
- The use of graphics to examine data is called ***visualization***.

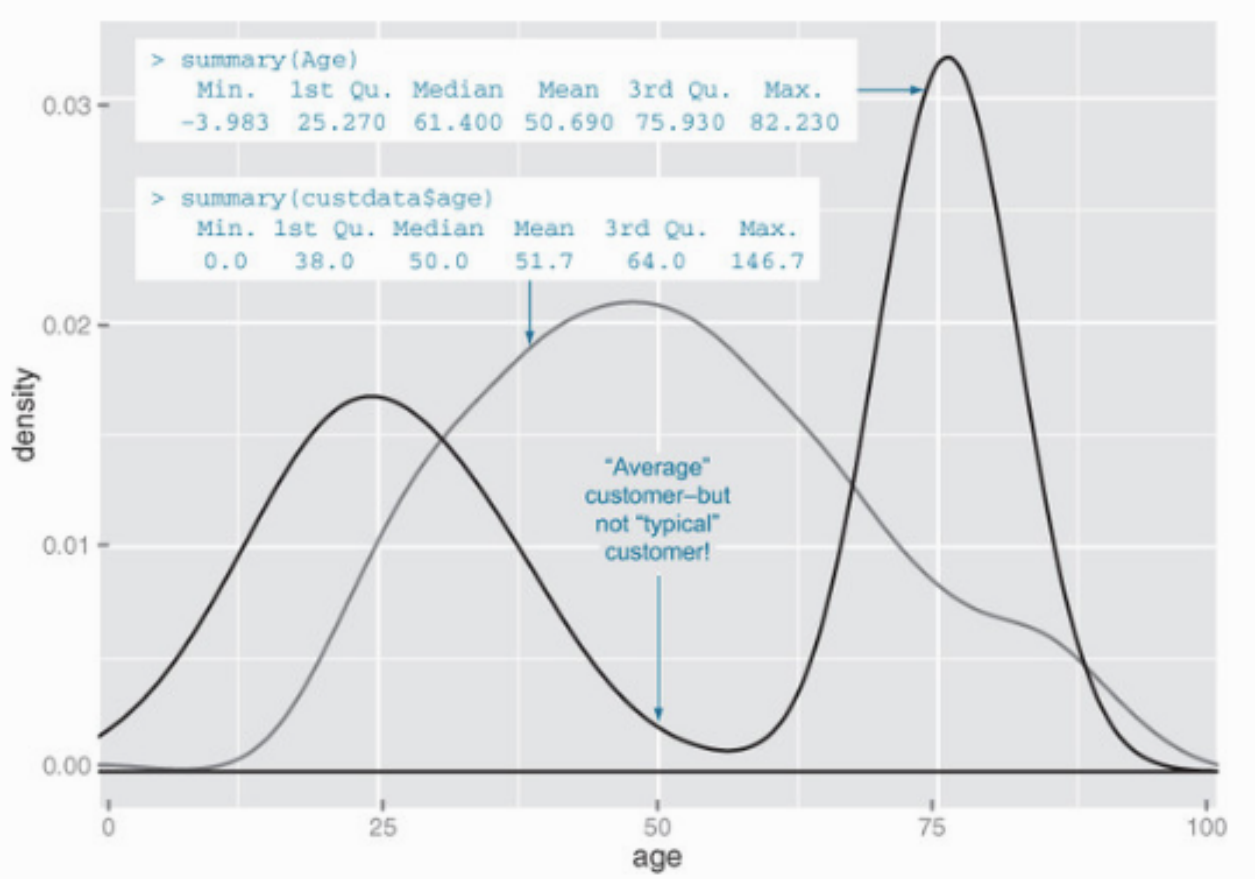
1. A graphic should display as much information as it can, with the lowest possible cognitive strain to the viewer.
2. Strive for clarity. Make the data stand out. Specific tips for increasing clarity include
   - Avoid too many superimposed elements, such as too many curves in the same graphing space.
   - Find the right aspect ratio and scaling to properly bring out the details of the data.
   - Avoid having the data all skewed to one side or the other of your graph.
3. Visualization is an iterative process. Its purpose is to answer questions about the data.

- During the visualization stage, you graph the data, learn what you can, and then re-graph the data to answer the questions that arise from your previous graphic.

Visually checking distributions for a single variable

- It helps answer the questions:
- What is the peak value of the distribution?
- How many peaks are there in the distribution (unimodality versus bimodality)?
- How normal (or lognormal) is the data?
- How much does the data vary? Is it concentrated in a certain interval or in a certain category?


The peak of the customer population is just under 50. That's not obvious from the summary.
> summary(custdata$age)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0 38.0 50.0 51.7 64.0 146.7
It's easier to read the mean, median and central 50% of the customer population off the summary.
It's easier to get a sense of the customer age range from the graph.
Customer "subpopulation": more customers over 75 than you would expect.
Invalid values?
Outliers
density
0.020
0.015
0.010
0.005
0.000
0
50
100
150
age


> summary(Age)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-3.983 25.270 61.400 50.690 75.930 82.230
> summary(custdata$age)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.0 38.0 50.0 51.7 64.0 146.7
"Average" customer–but not "typical" customer!
density
0.03
0.02
0.01
0.00
0
25
50
75
100
age

Histograms

- A basic histogram bins a variable into fixed-width buckets and returns the number of data points that falls into each bucket.

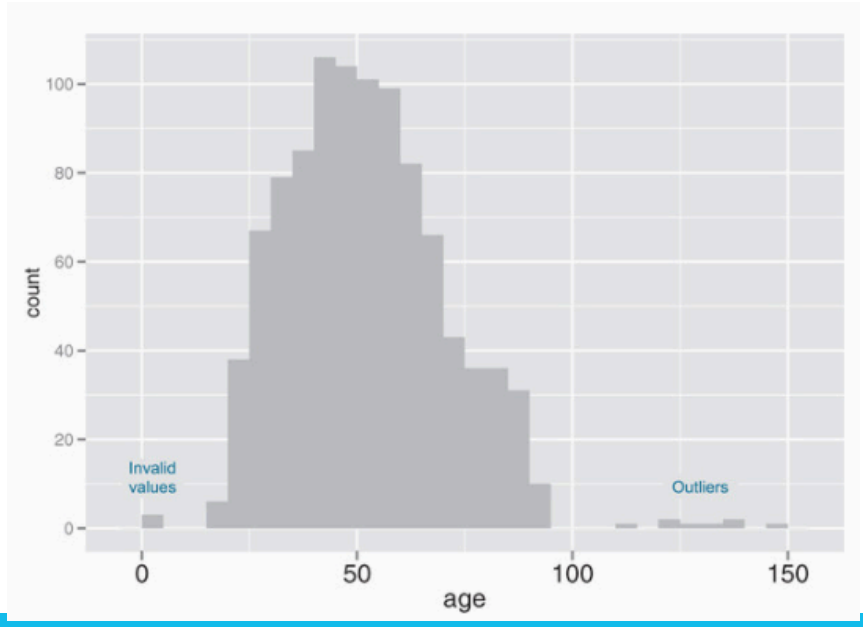
```
library(ggplot2)

ggplot(custdata) +
   geom_histogram(aes(x=age),
   binwidth=5, fill="gray")
```

Load the ggplot2 library, if you haven't already done so.

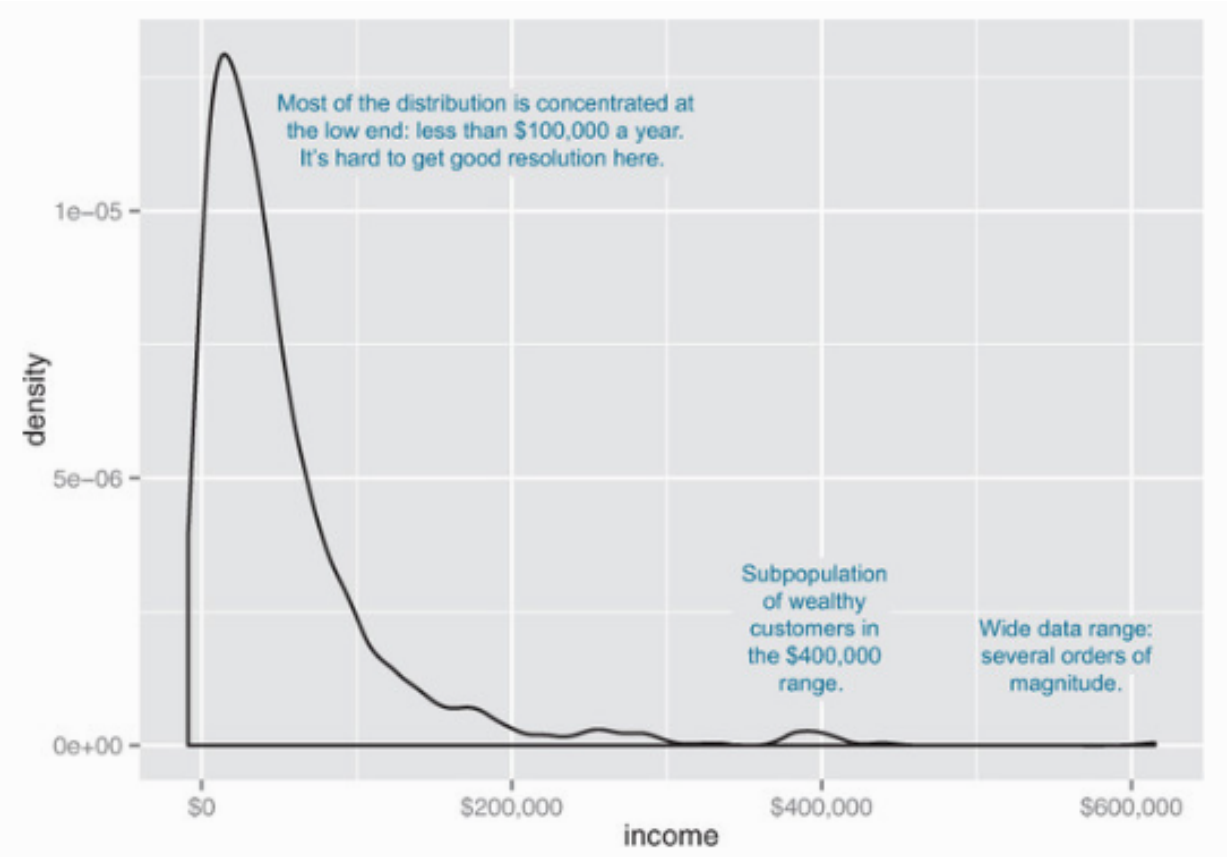
The binwidth parameter tells the geom_histogram call how to make bins of five-year intervals (default is datarange/30). The fill parameter specifies the color of the histogram bars (default: black).

- The primary disadvantage of histograms is that you must decide ahead of time how wide the buckets are.
- If the buckets are too wide, you can lose information about the shape of the distribution. If the buckets are too narrow, the histogram can look too noisy to read easily. An alternative visualization is the density plot.

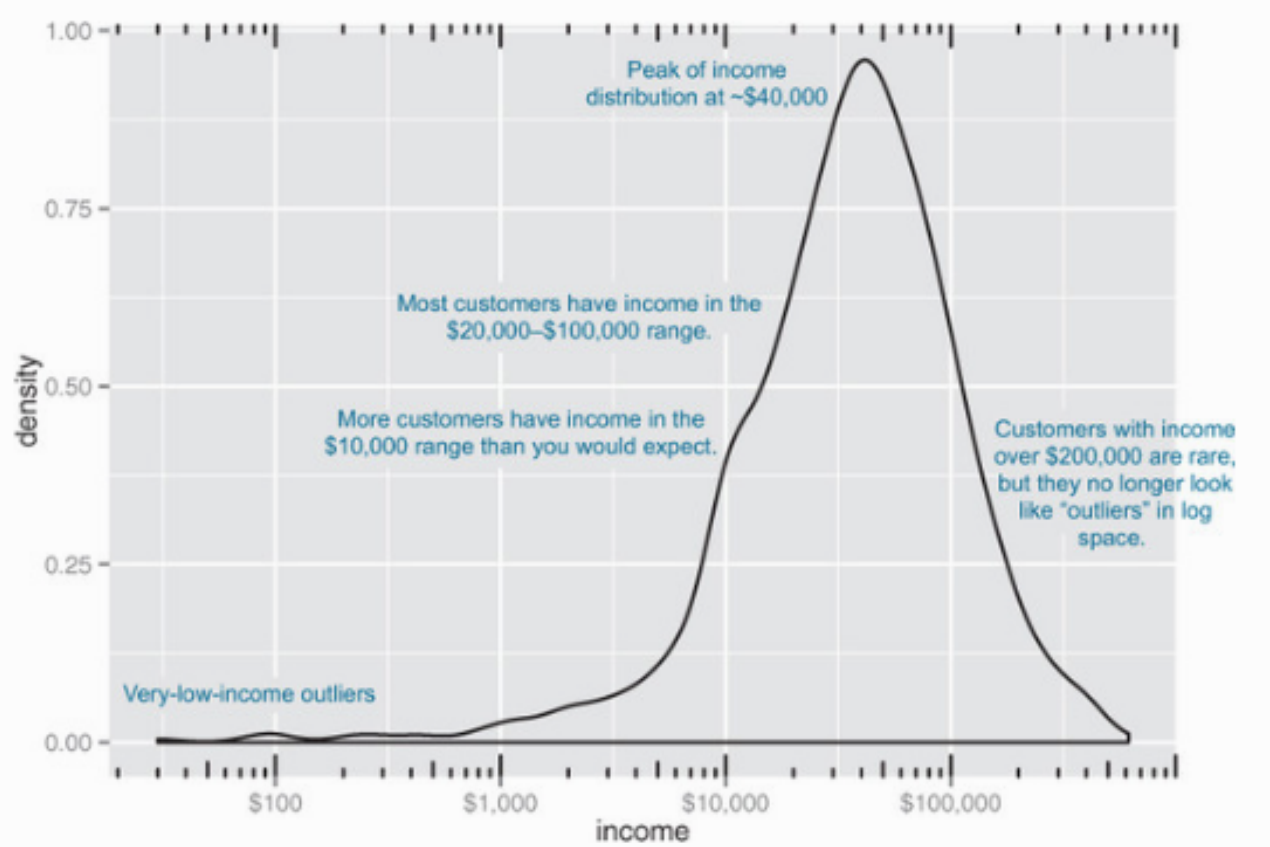


Density Plots

- Density plot is like continuous histogram" of a variable, except the area under the density plot is equal to 1. A point on a density plot corresponds to the fraction of data (or the percentage of data, divided by 100) that takes on a particular value. This fraction is usually very small
- density plot, you're more interested in the overall shape of the curve than in the actual values on the y-axis.
- When the data range is very wide and the mass of the distribution is heavily concentrated to one side, it's difficult to see the details of its shape.
- If the data is non-negative, then one way to bring out more detail is to plot the distribution on a logarithmic scale.



Density plots show where data is concentrated. This plot also highlights a population of higher-income customers.



The density plot of income on a log10 scale highlights details of the income distribution that are harder to see in a regular density plot.

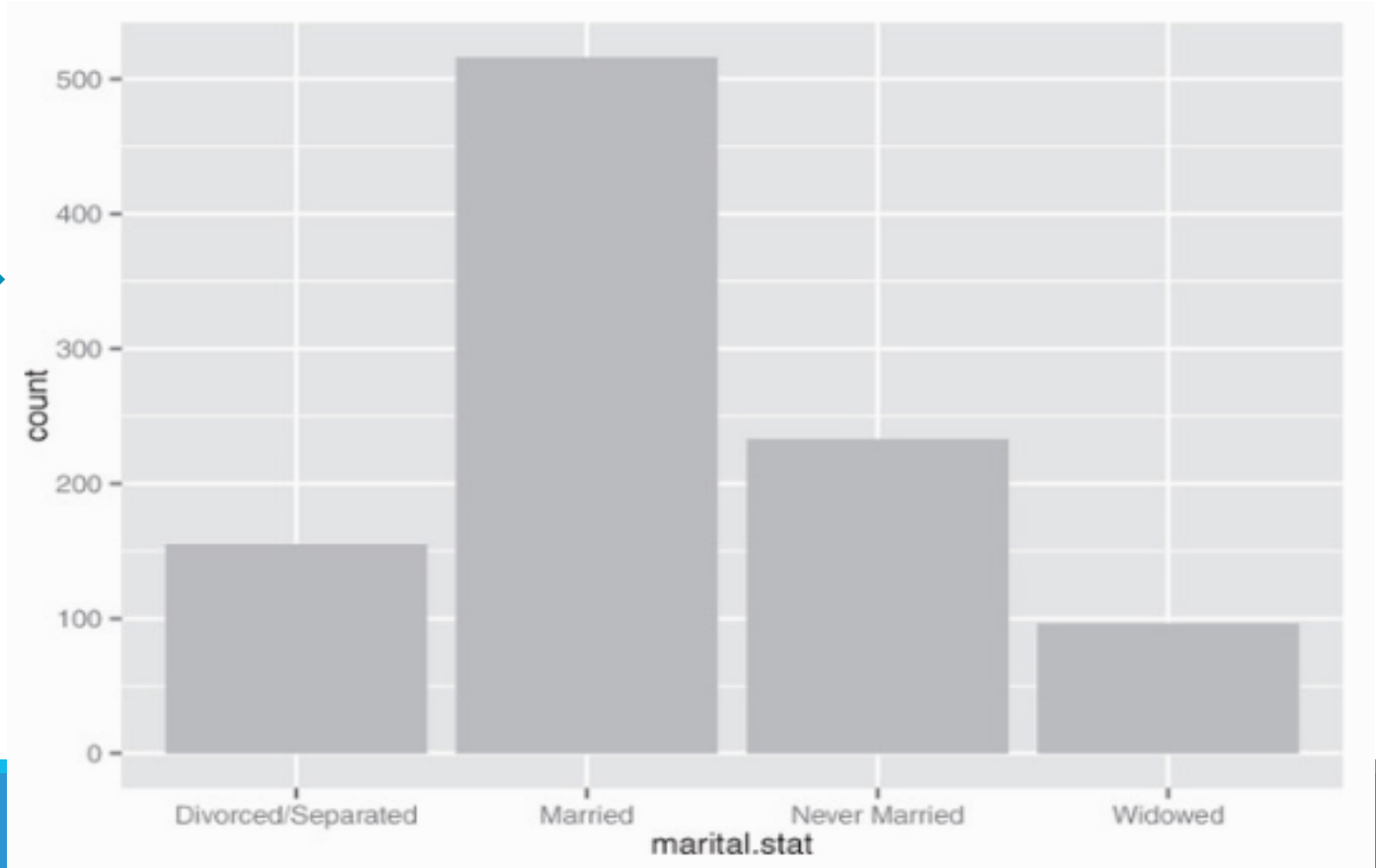
When should you use a logarithmic scale?

- Use a logarithmic scale when percent change, or change in orders of magnitude, is more important than changes in absolute units.
- Use a log scale to better visualize data that is heavily skewed.

Bar charts

A bar chart is a histogram for discrete data: it records the frequency of every value of a categorical variable.

Bar charts show the distribution of categorical variables.



Summary of Visualizations for one variable

Graph type	Uses
Histogram or density plot	Examines data range Checks number of modes Checks if distribution is normal/lognormal Checks for anomalies and outliers
Bar chart	Compares relative or absolute frequencies of the values of a categorical variable

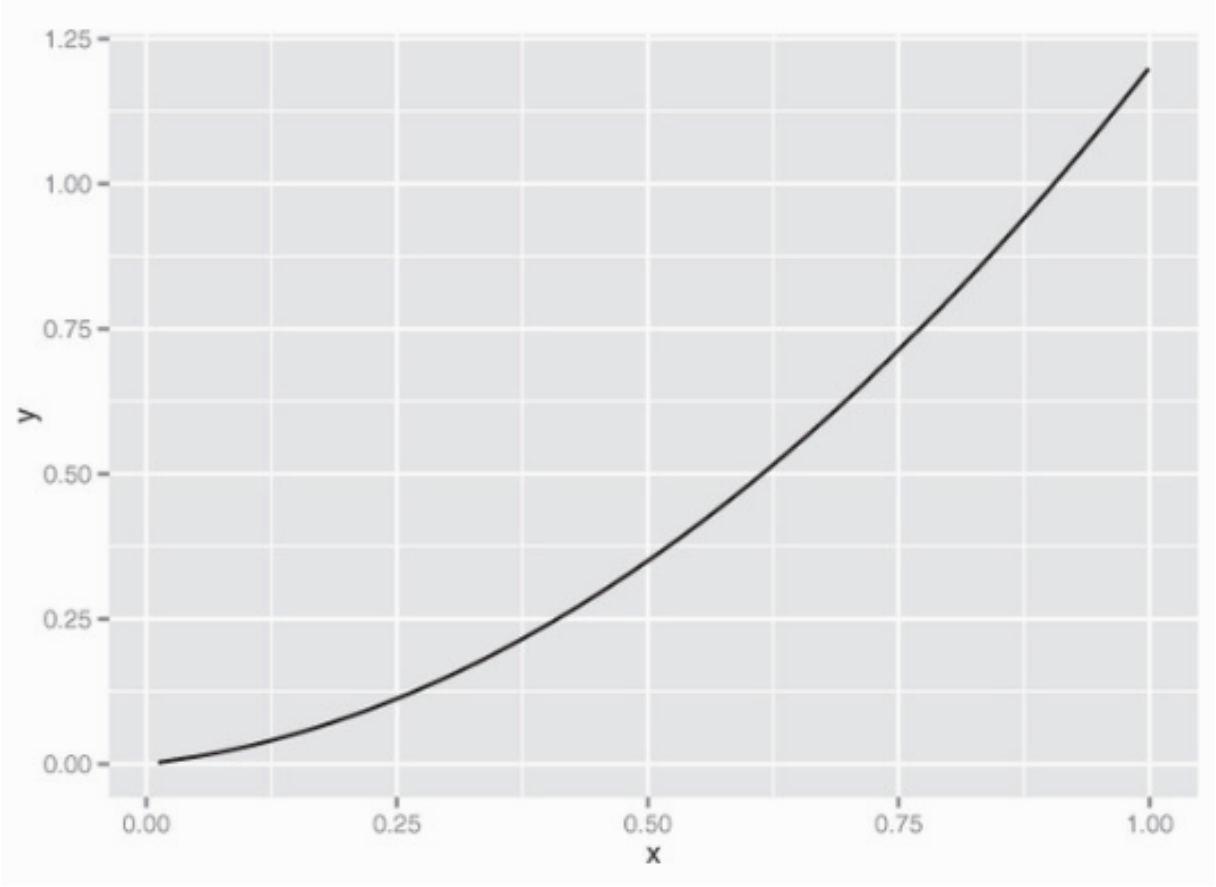
Visually checking relationships between two variables

- **You might want to answer questions like these:**

- Is there a relationship between the two inputs my data?
- What kind of relationship, and how strong?
- Is there a relationship between input 1 and the output of data? How strong?

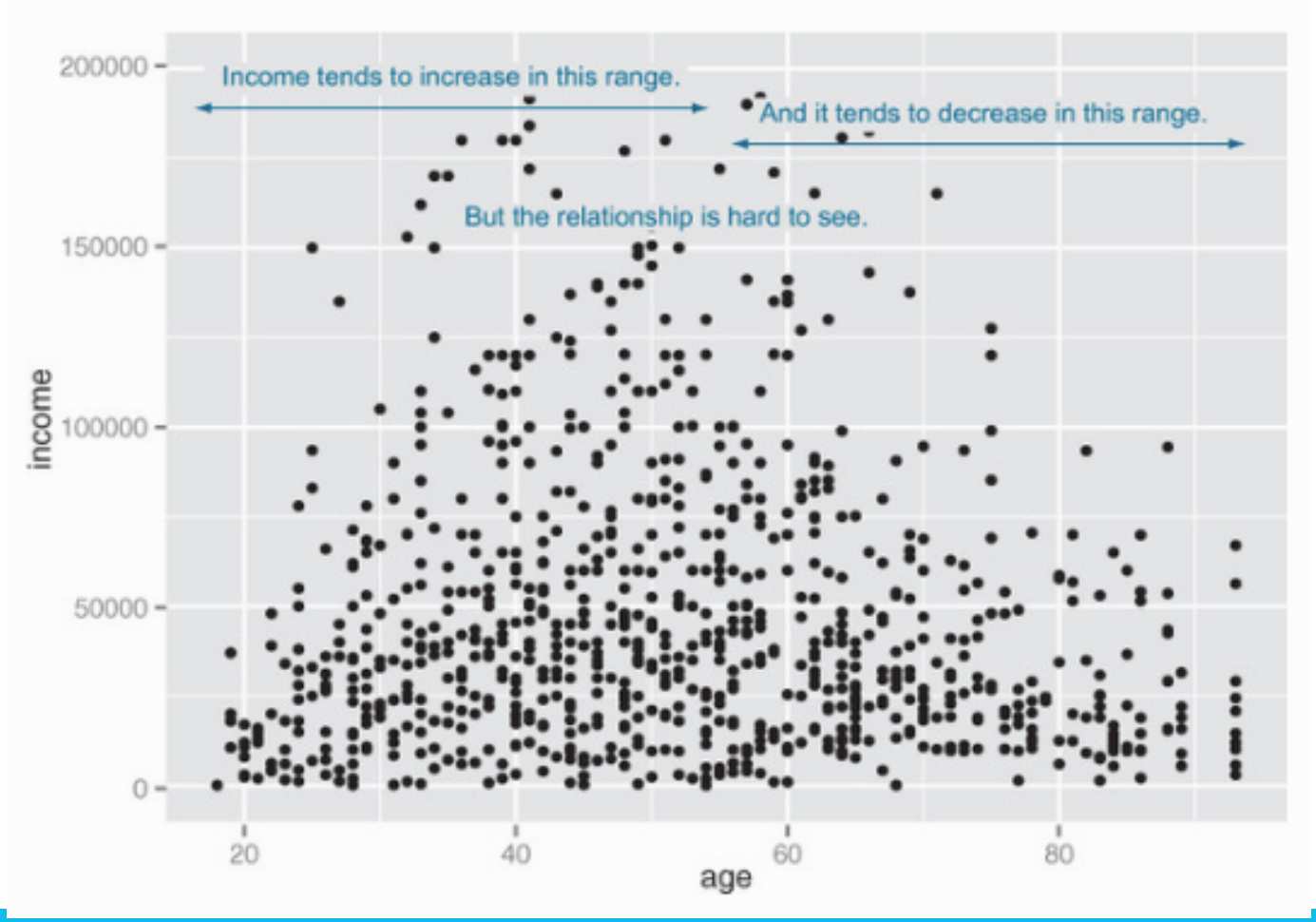
Line Plots

- Line plots work best when the relationship between two variables is relatively clean: each x value has a unique (or nearly unique) y value.

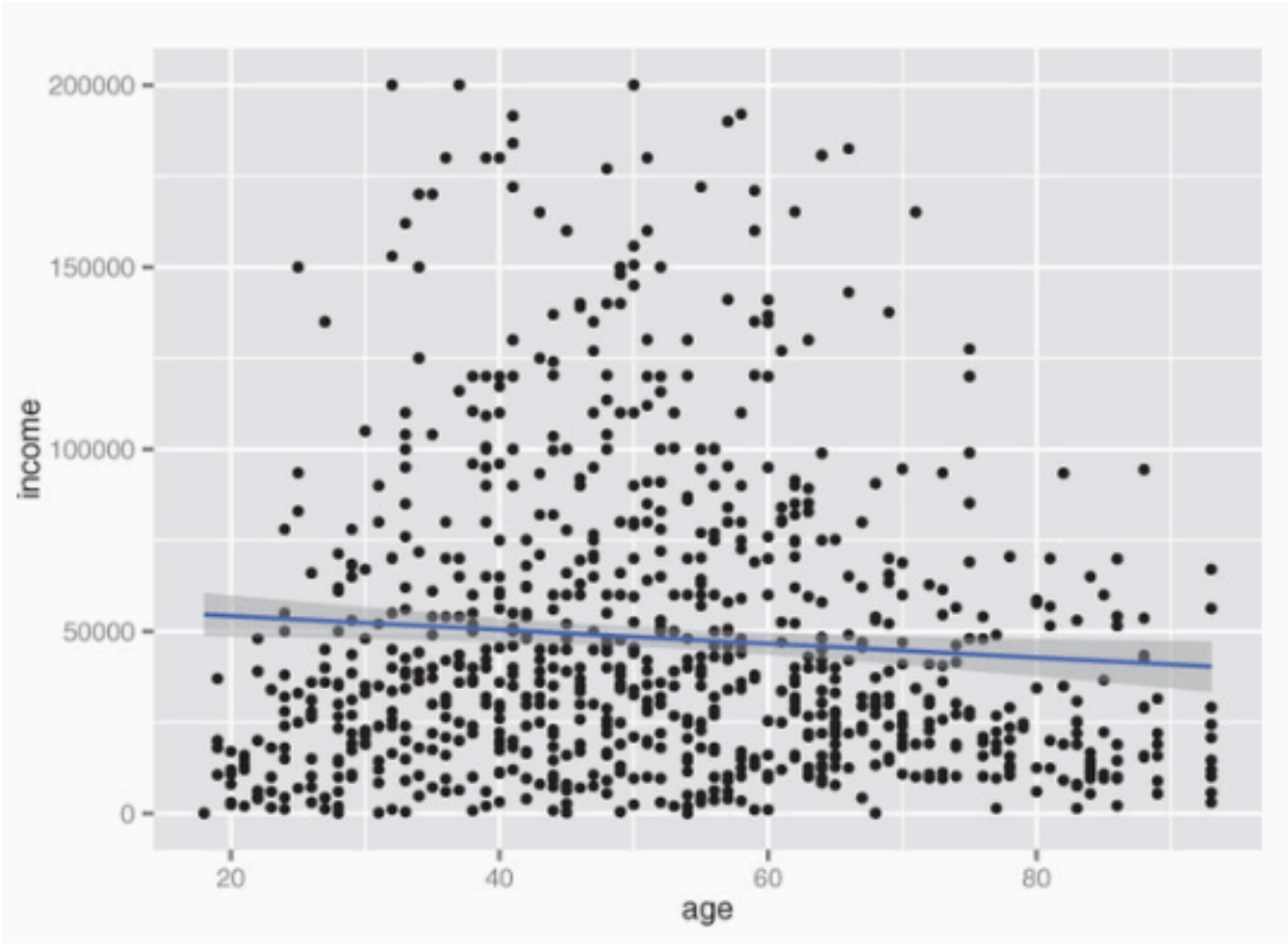


Scatter Plots

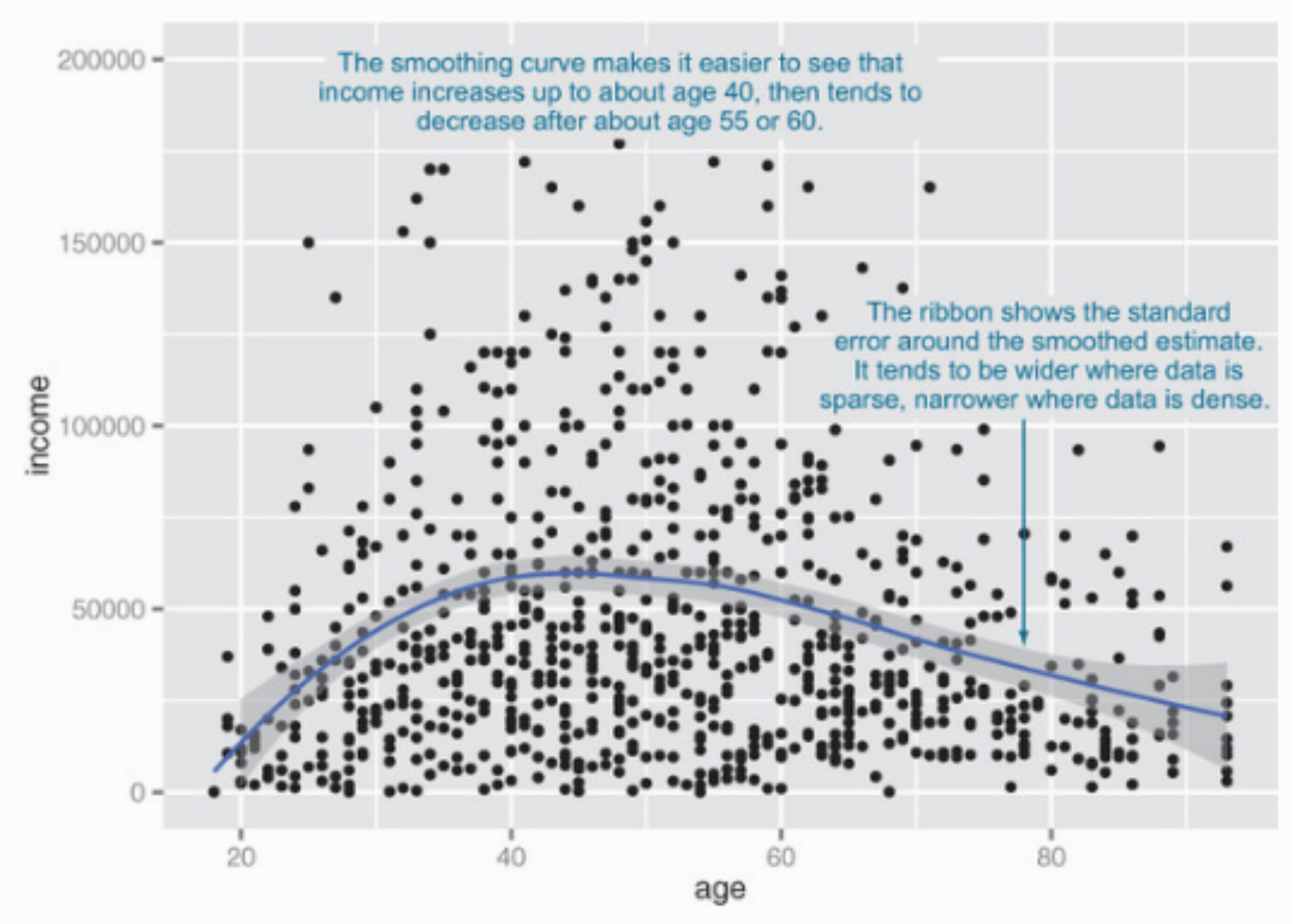
- When the data is not so cleanly related, line plots aren't as useful; you'll want to use the scatter plot instead.



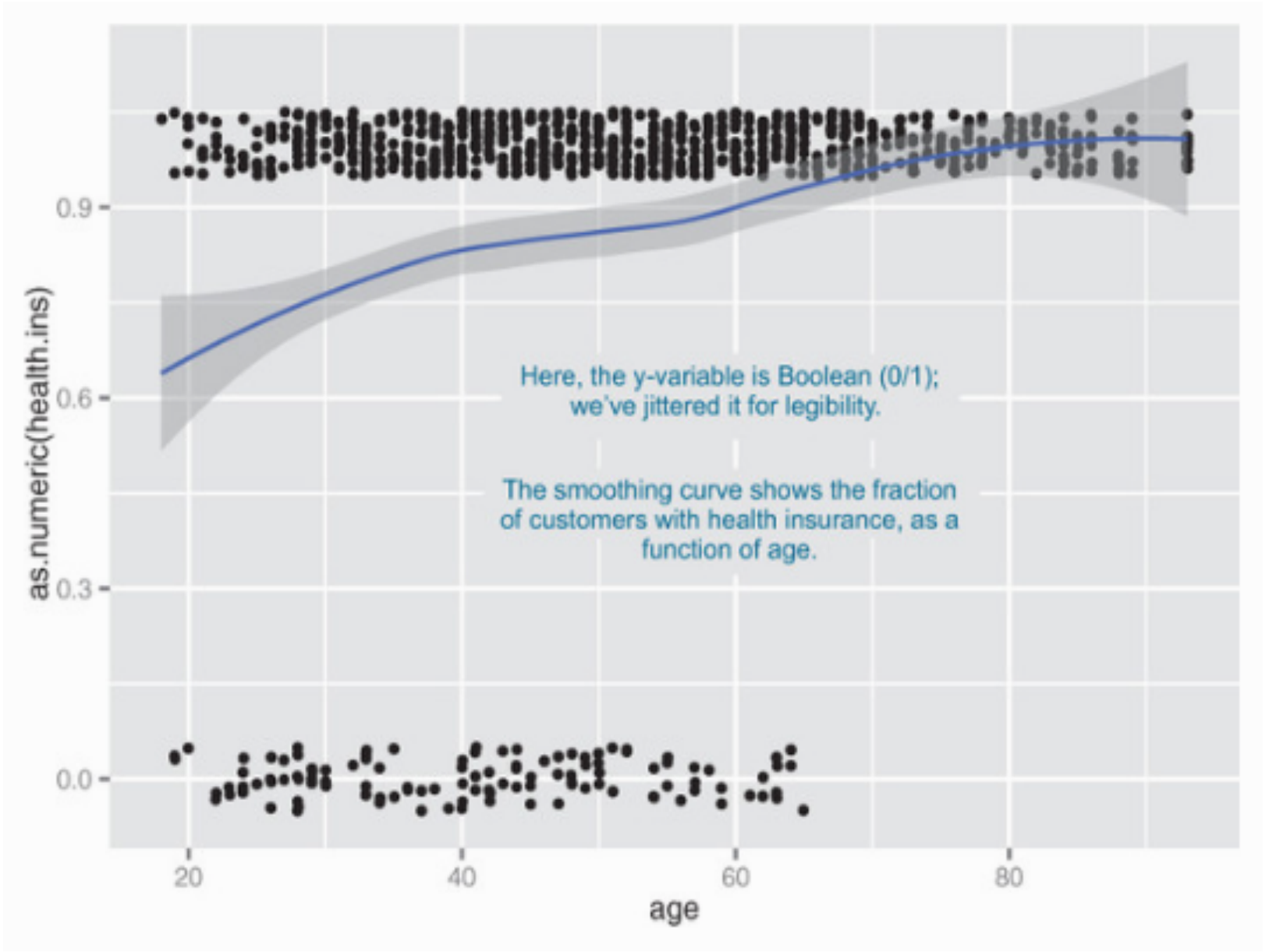
The relationship between age and income isn’t easy to see in the previous figure. Therefore, you can try to make the relationship clearer by also plotting a linear fit through the data.



The linear fit doesn't really capture the shape of the data. You can better capture the shape by instead plotting a smoothing curve through the data.

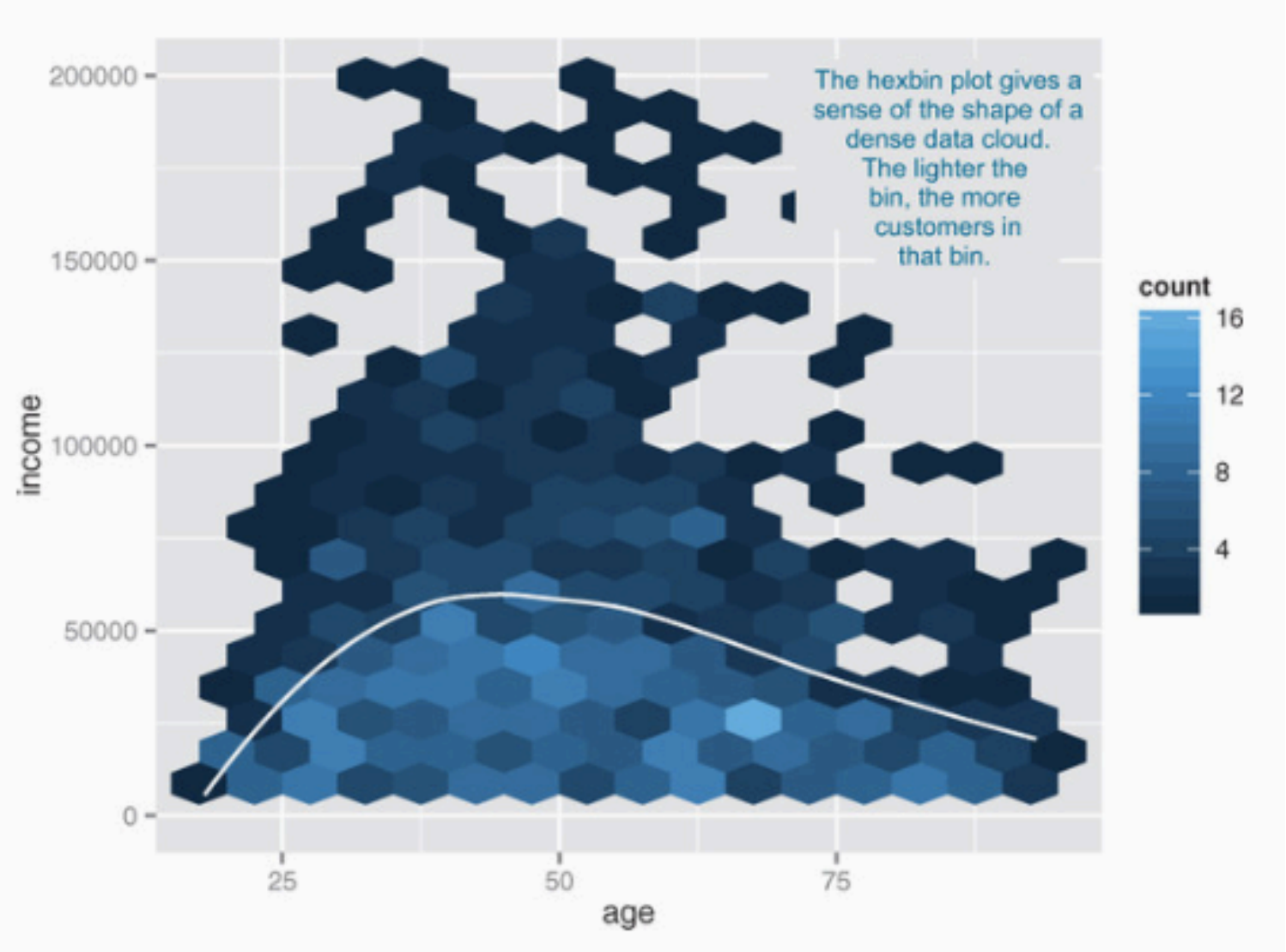


- A scatter plot with a smoothing curve also makes a good visualization of the relationship between a continuous variable and a Boolean.
- This show the probability of having health insurance increases as customer age increases.



Hexbin plot

- f the dataset were a hundred times bigger, there would be so many points that they would begin to plot on top of each other; the scatter plot would turn into an illegible smear.
- In high-volume situations like this, try an aggregated plot, like a hexbin plot.
- Ahexbin plot is like a two-dimensional histogram. The data is divided into bins, and the number of data points in each bin is represented by color or shading.

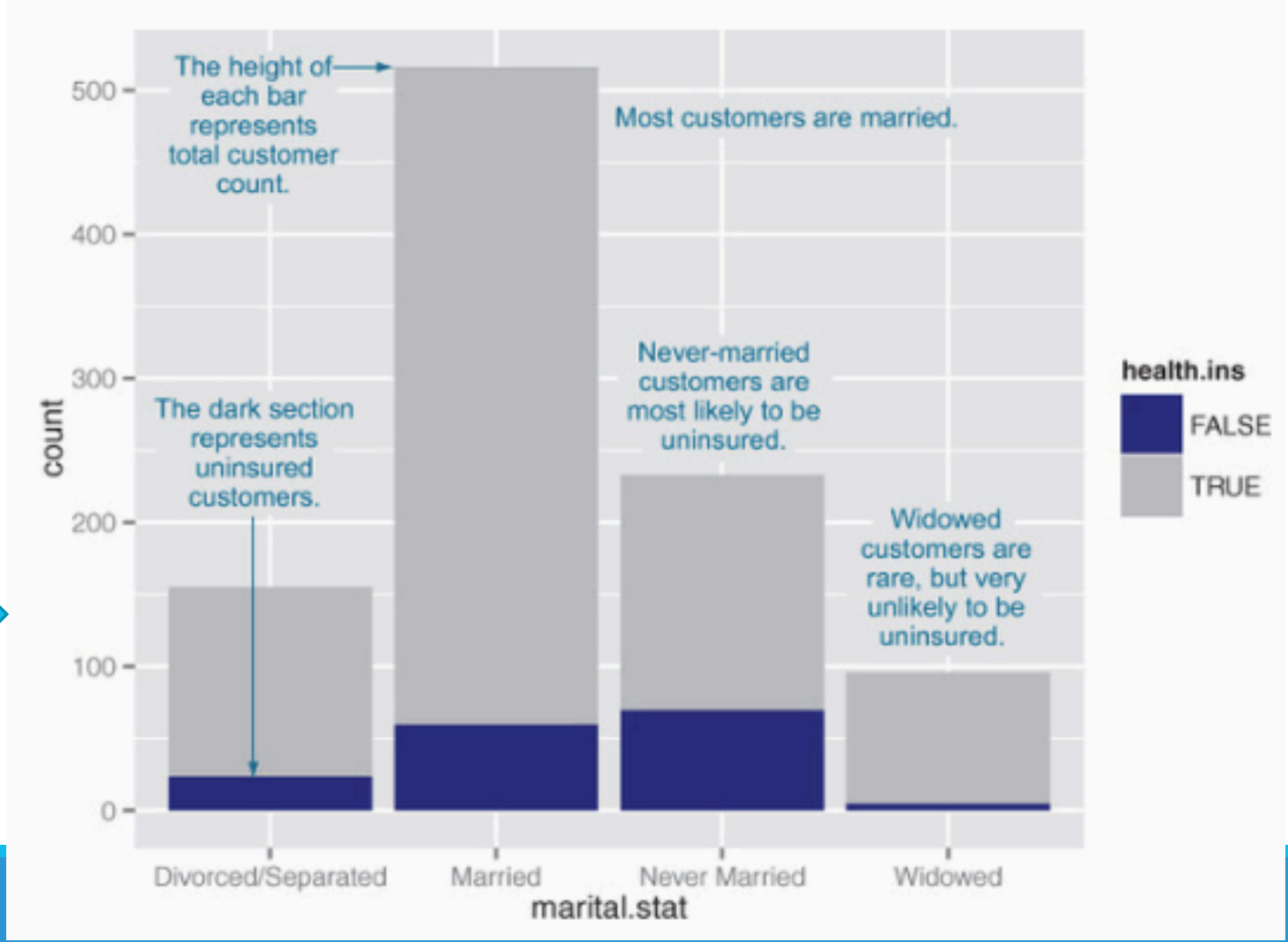


Hexbin plot of income versus age, with a smoothing curve superimposed in white.

Bar charts for two categorical variables

- ***stacked bar*** chart is straightforward way to examine relationship between two categorical variables.

Health insurance versus marital status



Summarizes the visualizations for two variables that we've covered.

Graph Type	Uses
Line Plot	Shows the relationship between two continuous variables. Best when that relationship is functional, or nearly so.
Scatter Plot	hows the relationship between two continuous variables. Best when the relationship is too loose or cloud-like to be easily seen on a line plot.
Smoothing Curve	Shows underlying average" relationship, or trend, between two continuous variables. Can also be used to show the relationship between a continuous and a binary or Boolean variable: the fraction of true values of the discrete variable as a function of the continuous variable.
Hexbin Plot	Shows the relationship between two continuous variables when the data is very dense.
Stacked bar chart	Shows the relationship between two categorical variables (var1 and var2). Highlights the frequencies of each value of var1.

Summarizes the visualizations for two variables -Cont.

Graph Type	Uses
Side-by-side bar chart	Shows the relationship between two categorical variables (var1 and var2). Good for comparing the frequencies of each value of var2 across the values of var1. Works best when var2 is binary.
Filled bar chart	Shows the relationship between two categorical variables (var1 and var2). Good for comparing the relative frequencies of each value of var2 within each value of var1. Works best when var2 is binary.
Bar chart with faceting	Shows the relationship between two categorical variables (var1 and var2). Best for comparing the relative frequencies of each value of var2 within each value of var1 when var2 takes on more than two values.

Key Takeaways

- Take the time to examine your data before diving into the modeling.
- The `summary` command helps you spot issues with data range, units, data type, and missing or invalid values.
- Visualization additionally gives you a sense of data distribution and relationships among variables.
- Visualization is an iterative process and helps answer questions about the data. Time spent here is time not wasted during the modeling process.