Week 1, January 13th 2017

General structure

- First 30 mins: recap / review / questions
- Second 90 mins: new material lecture
- Third 60 mins: in-class work on data analysis
- I have to leave at 12:00 on the nose
- Breaks?

Evaluation

- Five assignments (5% - 25% graded by me and/or Jessie)
- Group presentation (5% graded by the class)
- In-class, closed-book midterm (20% graded by me)
- I expect you to strive for excellence
- I am a HARD marker

Assignments/Midterm

- #1: 15%, questions and answers
- #2-#3: 5% each, 2-page reports
- #4: 25%, mock short scientific article
- #5a: 25%, sampling strategy group project
- #5b: 5%, sampling strategy group presentation
- Midterm: 20%, 90 minutes in-class

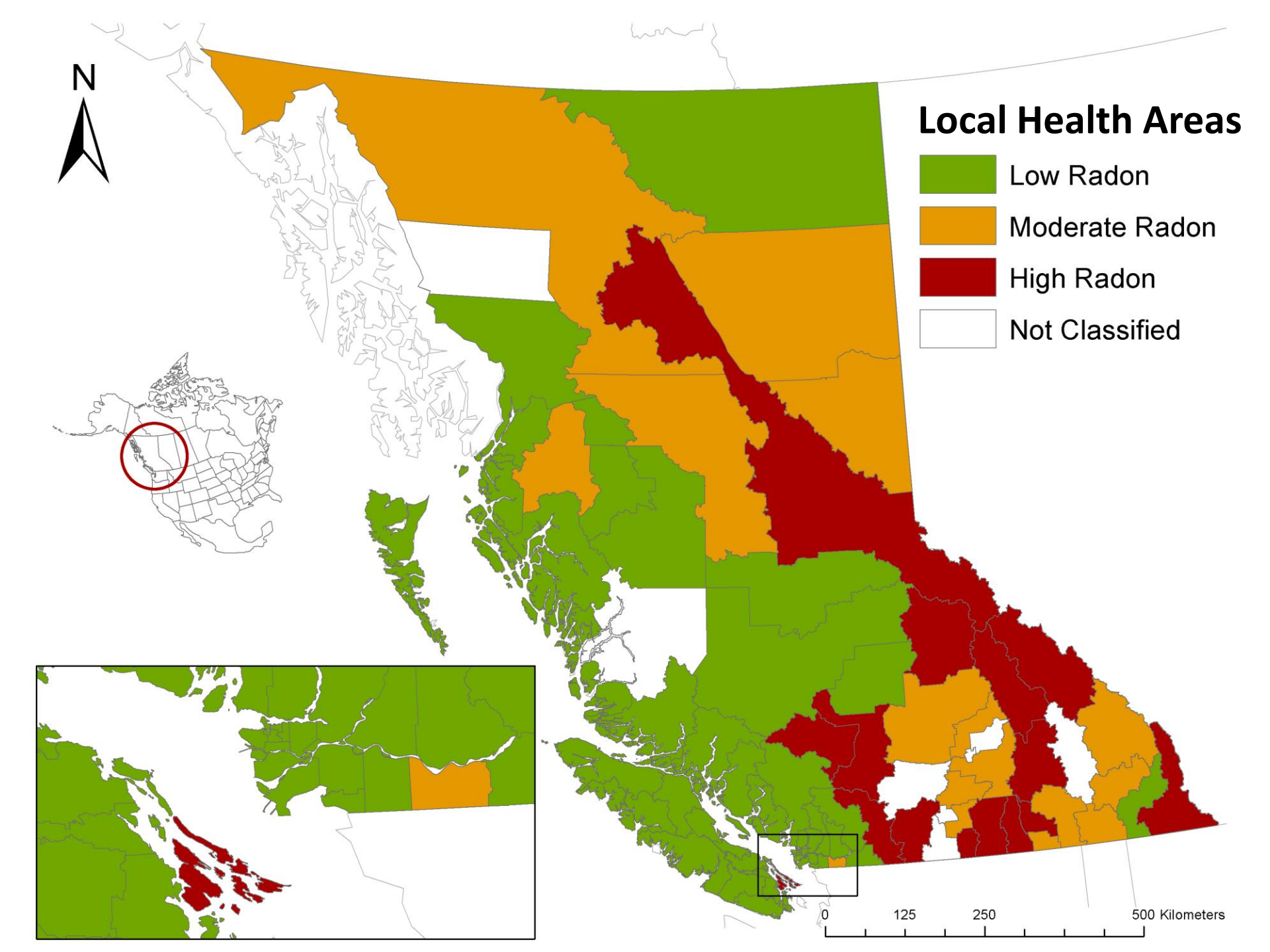
Communications

- At UBC on Fridays only, contact me to meet
- My office is 327B in SPPH
- Questions about course material, assignment details, etc. MUST be emailed to the entire class
- Phone calls should be minimal

Introduction to Radon

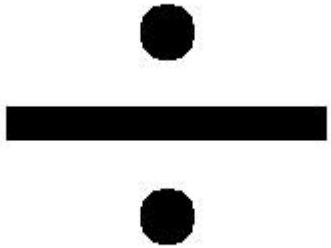
Health Canada says…

<200 Bq/m³	Below guideline
200 – 600 Bq/m³	Fix within **2** years
>600 Bq/m³	Fix within **1** year


N
Local Health Areas
Low Radon
Moderate Radon
High Radon
Not Classified
0
125
250
500 Kilometers

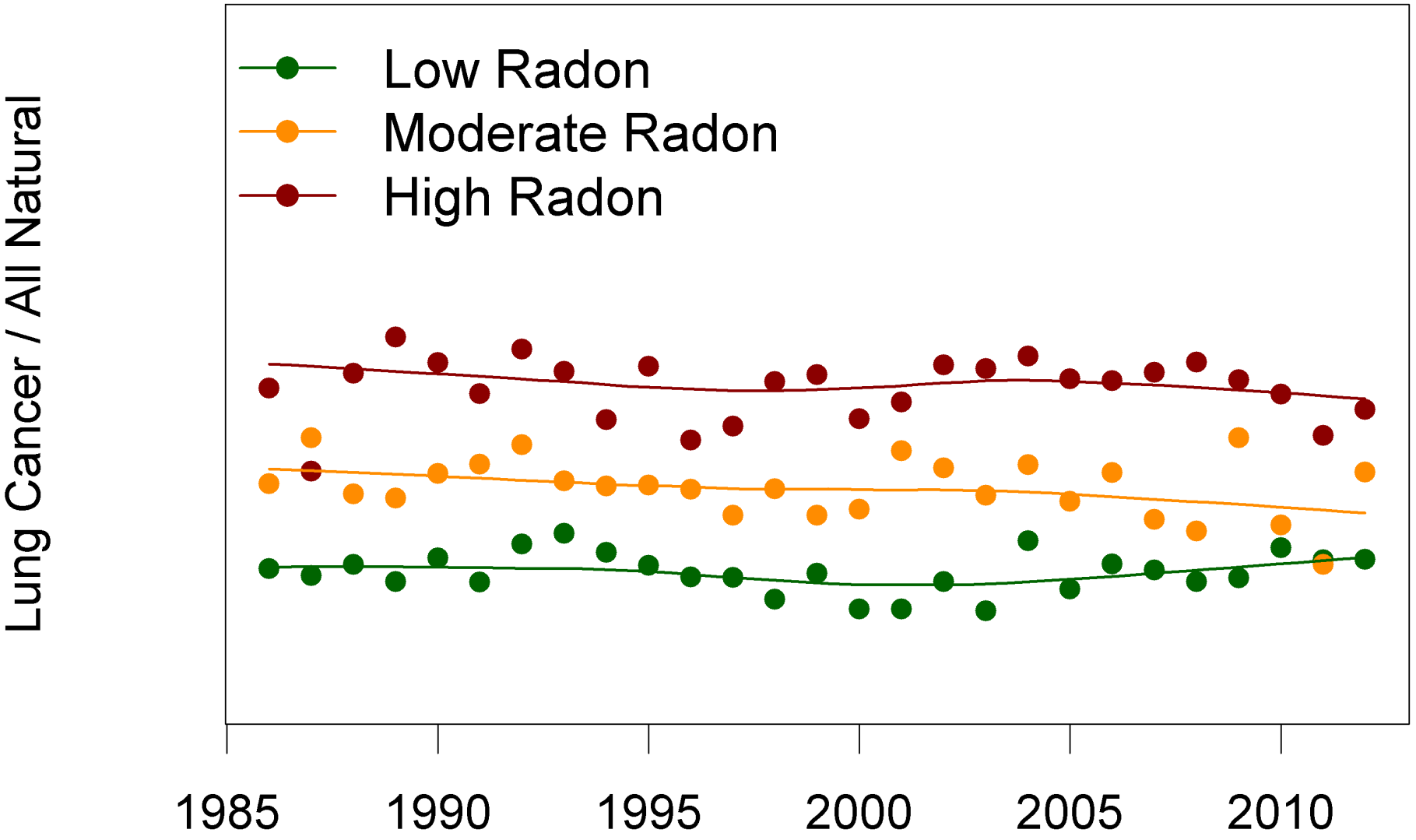
1986 – 2012 Trends

Number of annual deaths from lung cancer

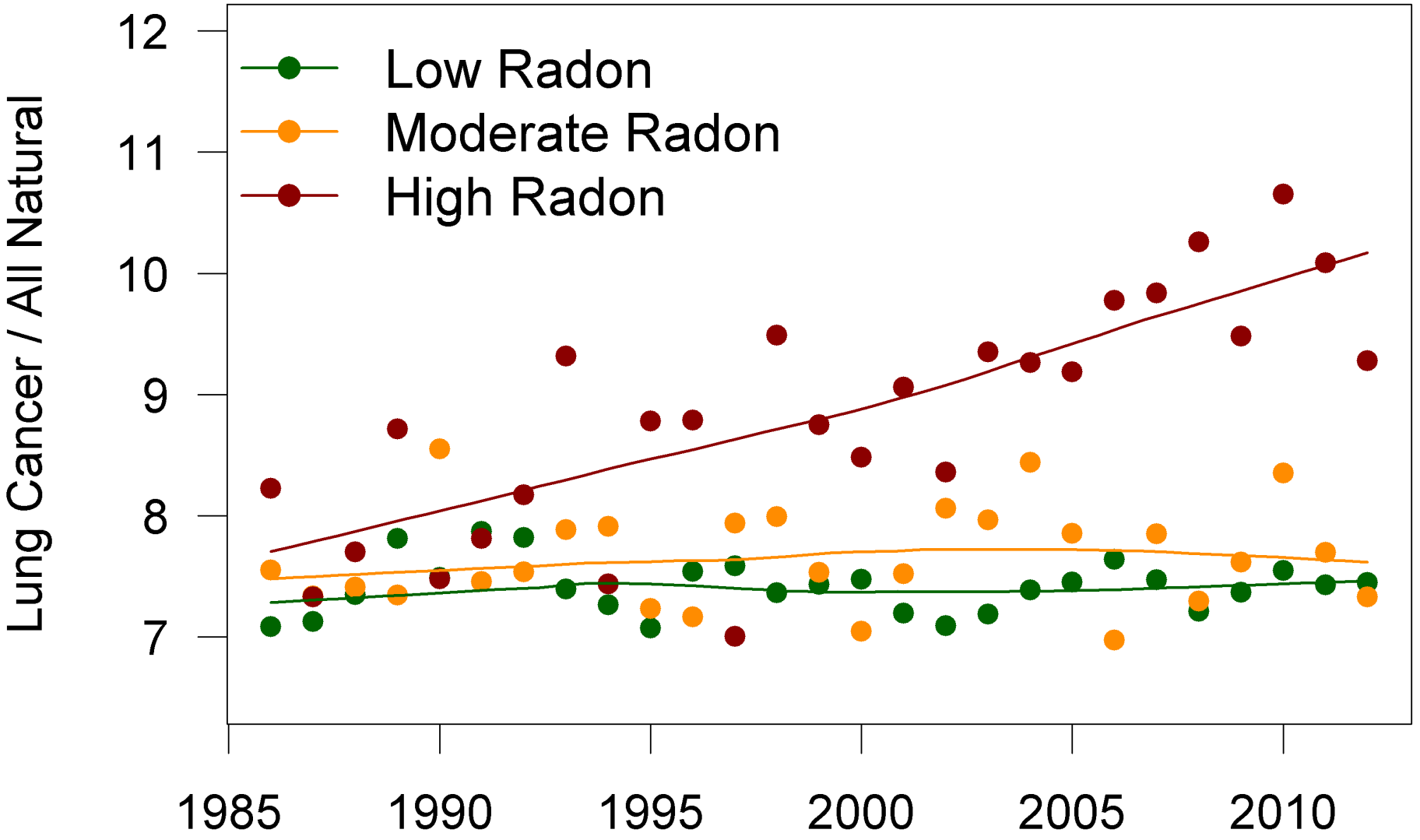


Number of annual deaths from all natural causes

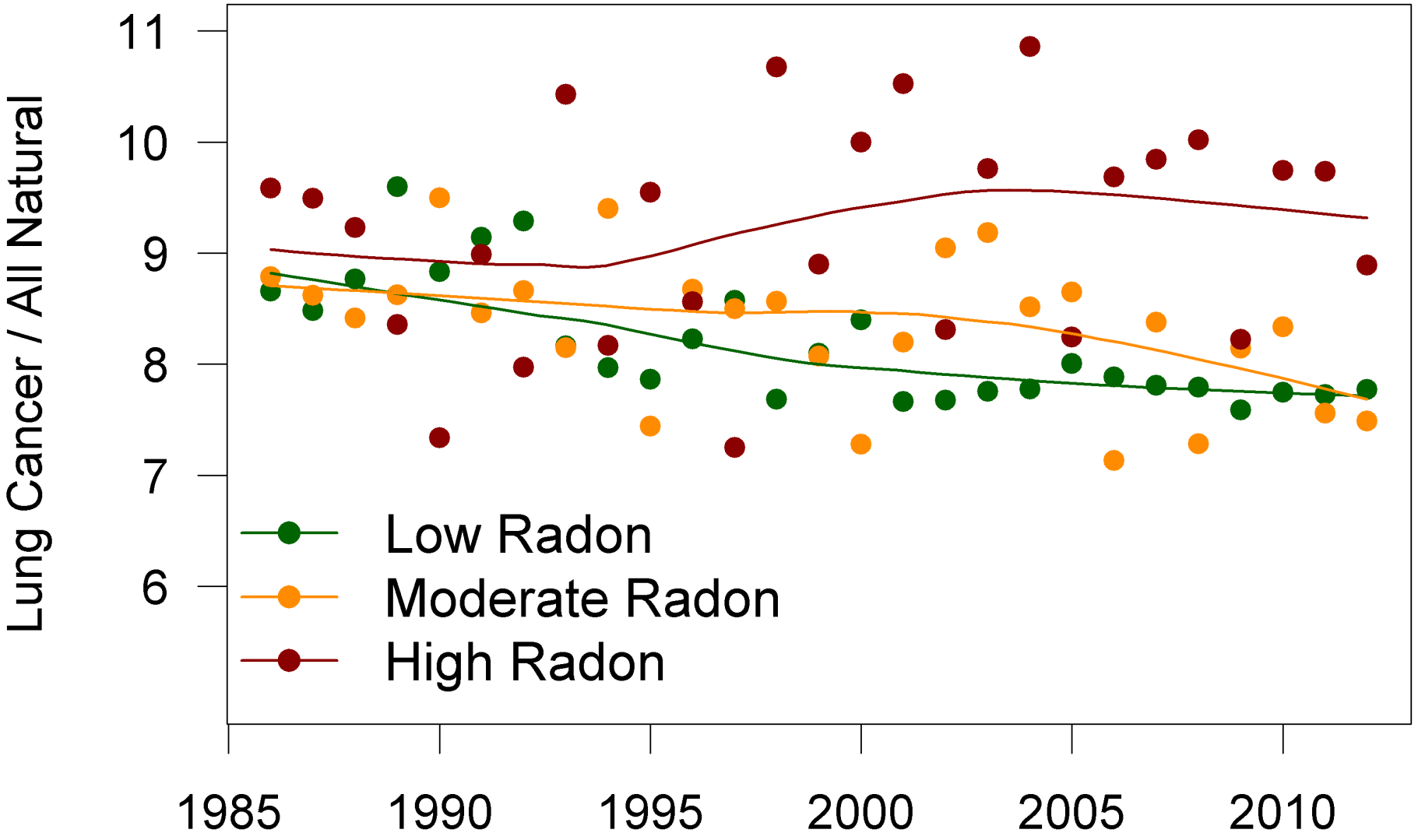
Hypothetical



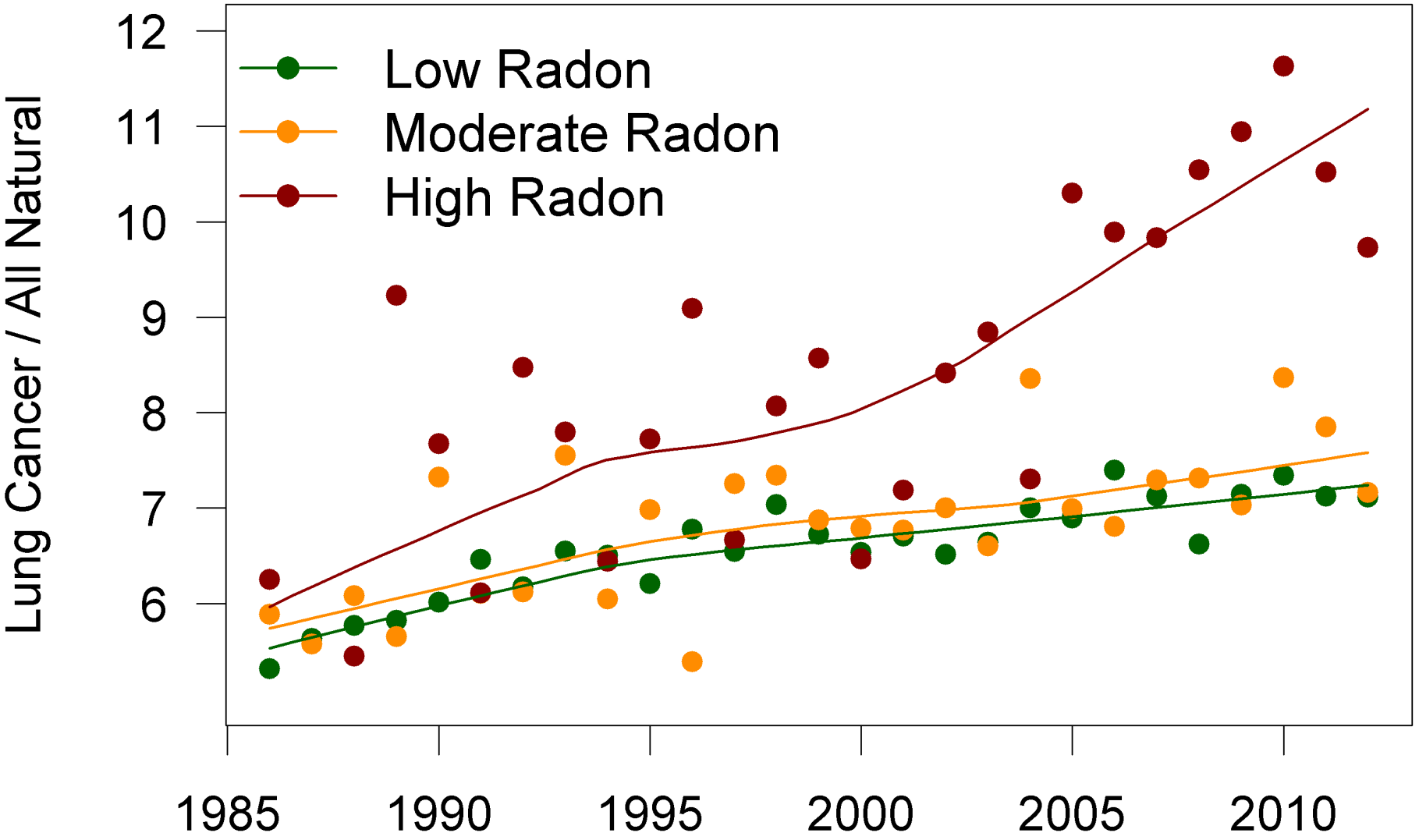
All deaths


Lung Cancer / All Natural
12
11
10
9
8
7
Low Radon
Moderate Radon
High Radon
1985
1990
1995
2000
2005
2010

Male

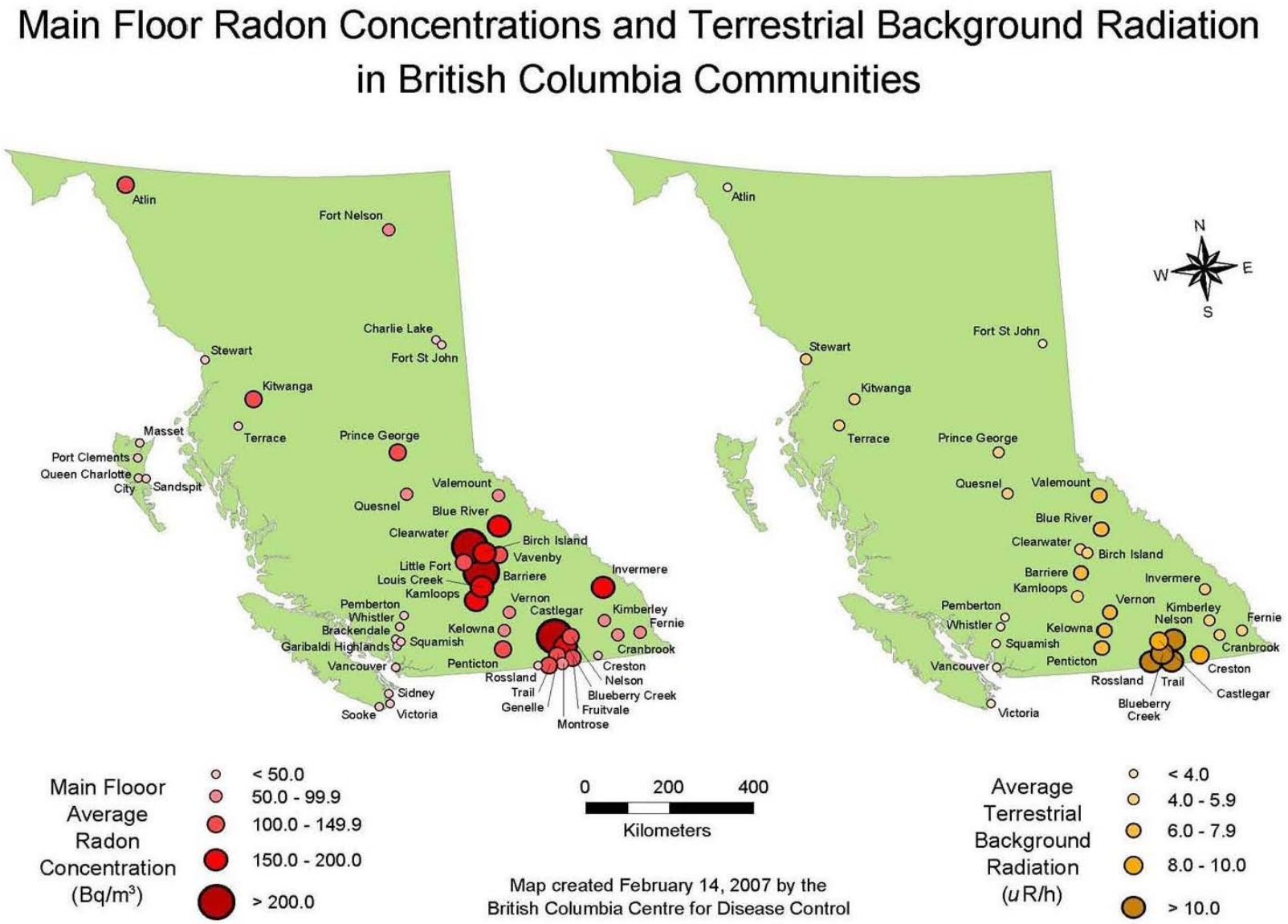

Lung Cancer / All Natural
11
10
9
8
7
6
1985
1990
1995
2000
2005
2010
Low Radon
Moderate Radon
High Radon

Female


Lung Cancer / All Natural
12
11
10
9
8
7
6
Low Radon
Moderate Radon
High Radon
1985
1990
1995
2000
2005
2010

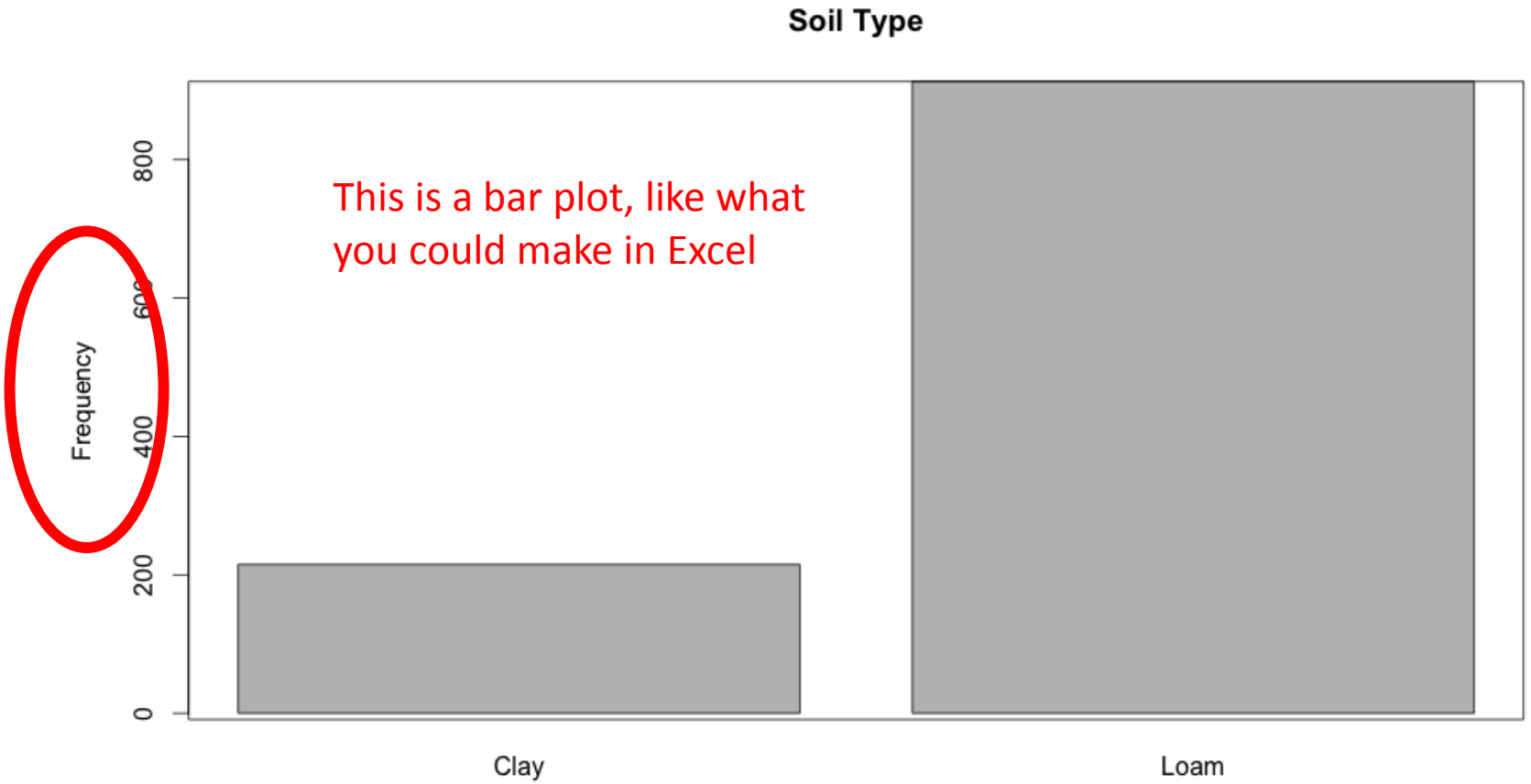
Question: can we predict high risk indoor radon concentrations using data about homes and where they are located?

BCCDC Radon Dataset



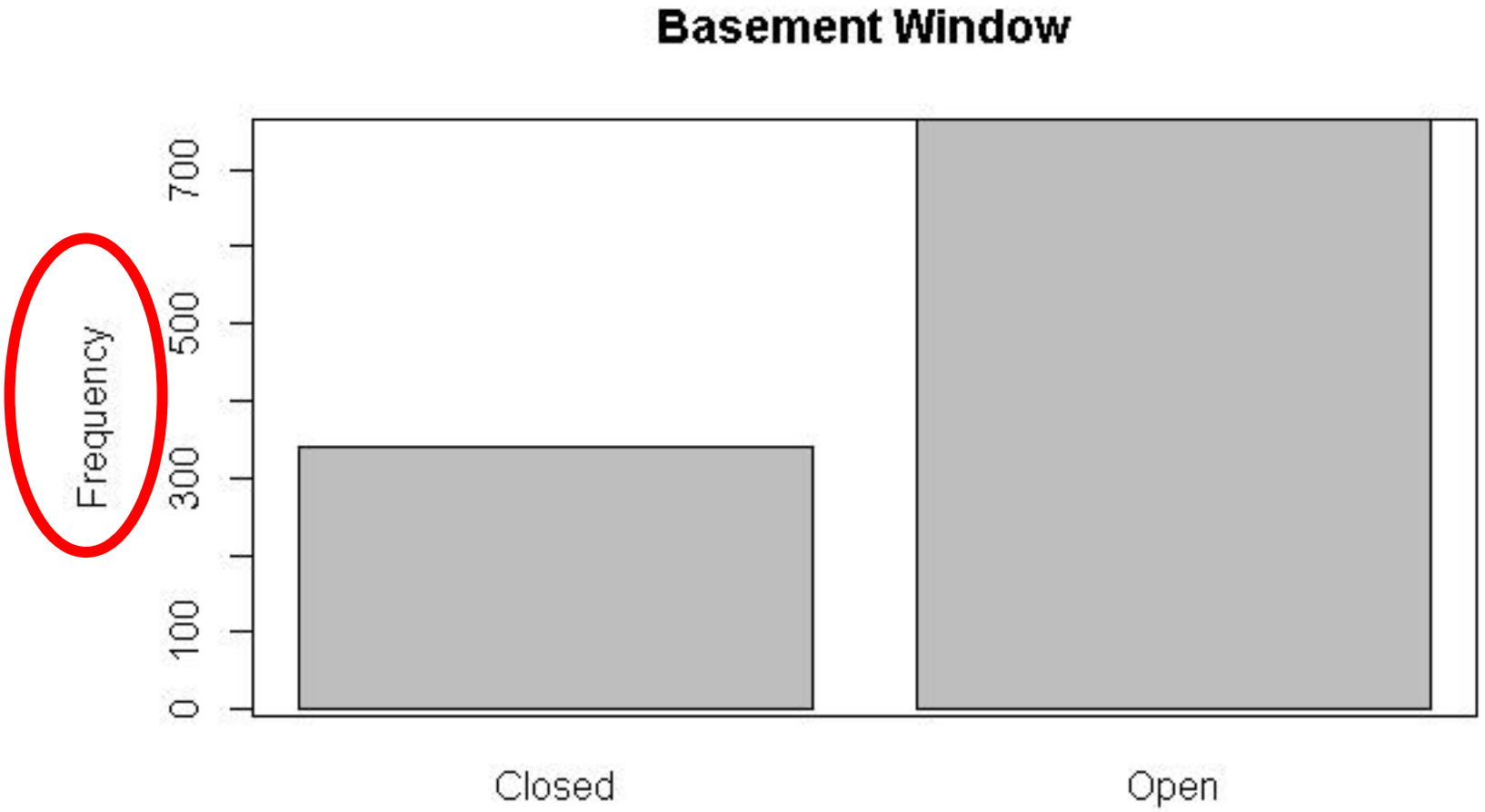
Dichotomous Variables

- Can take one of two values



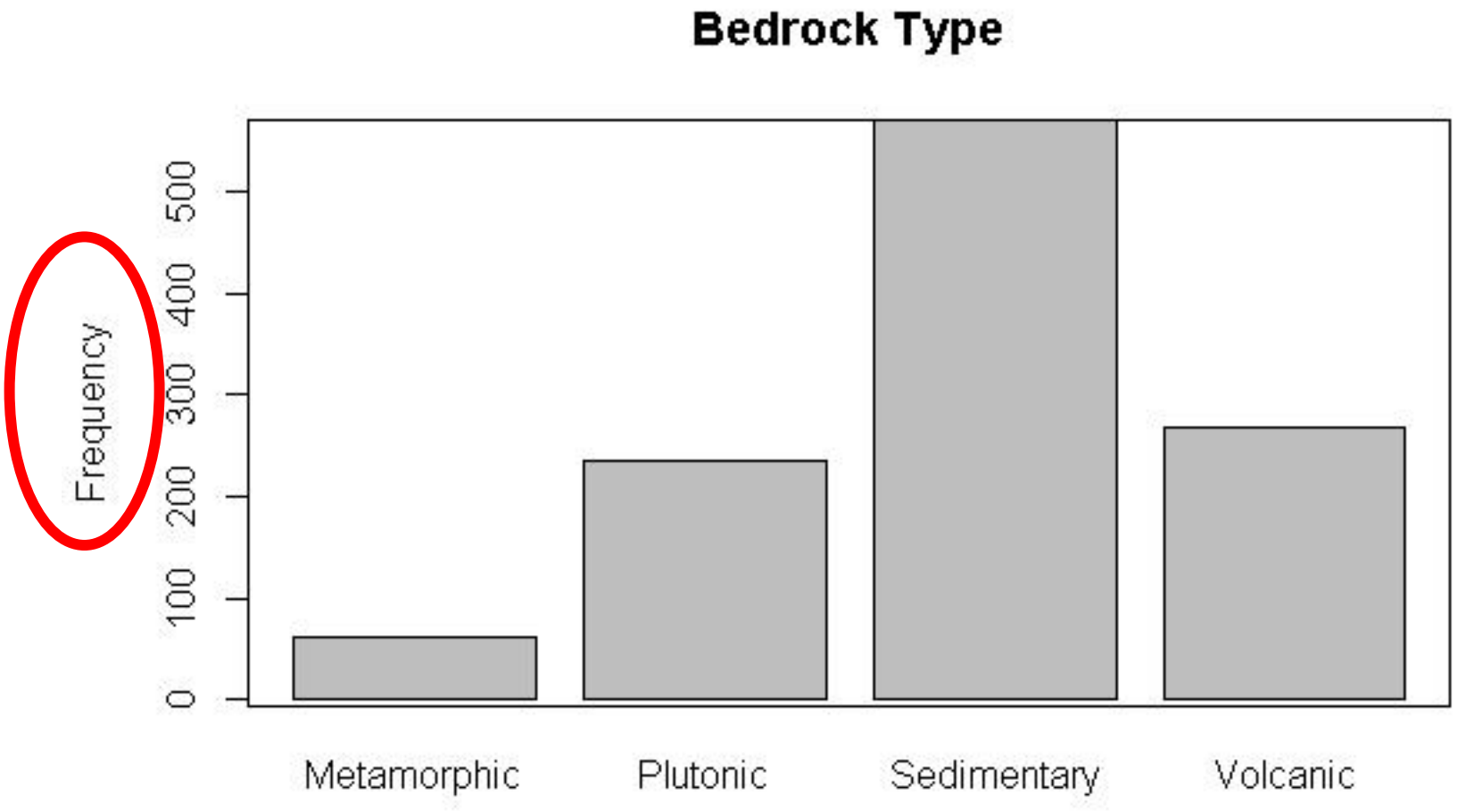
Binary Variables

- Specific type of dichotomous variable where the values are in opposition to each other (i.e. on/off, yes/no, open/closed)
- The terms ‘dichotomous’ and ‘binary’ are often used interchangeably, but it’s good to understand the technical distinction



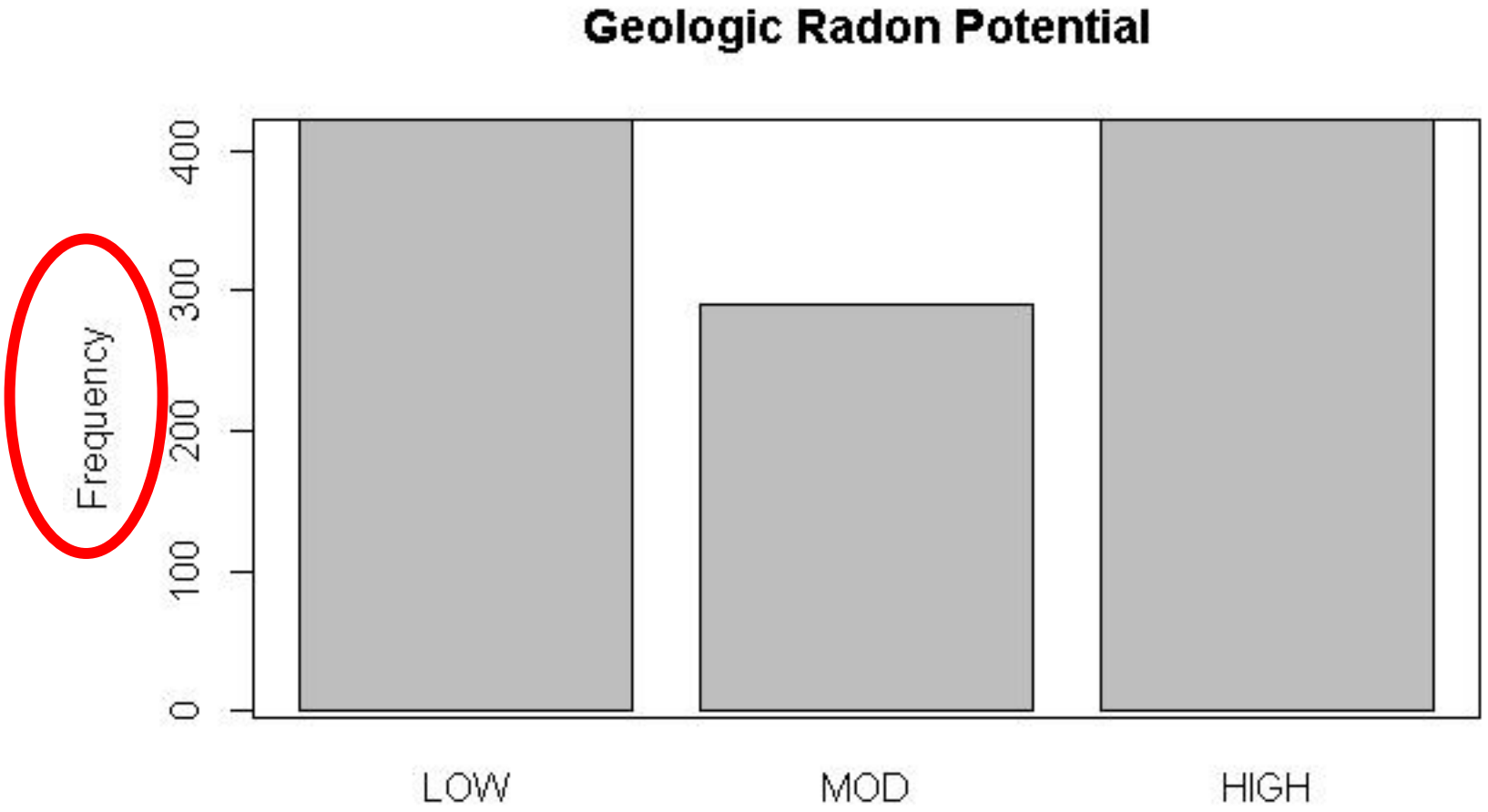
Categorical Variables

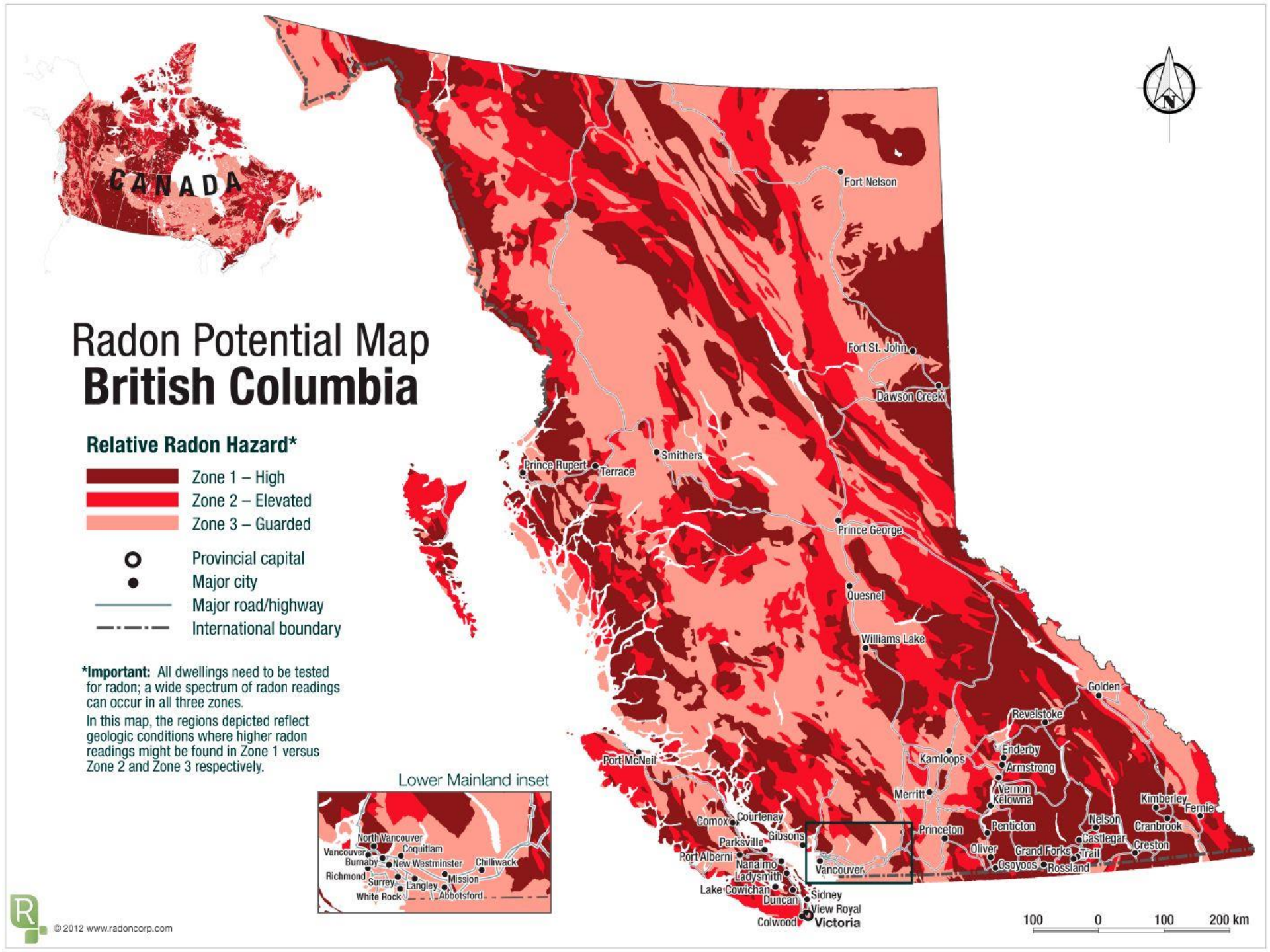
- Can take one of more than two values
- The number of possible values is limited and fixed



Ordinal Variables

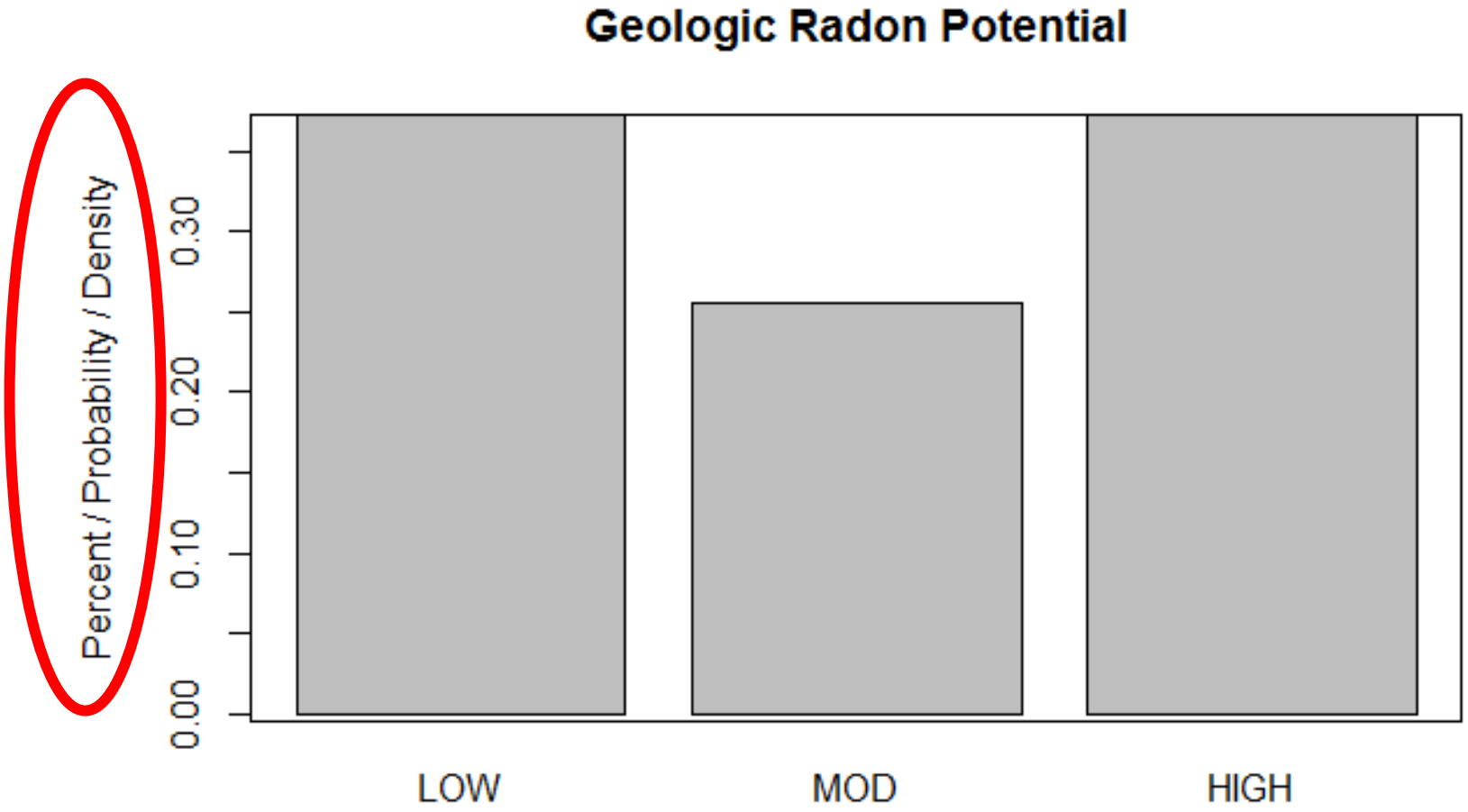
- Specific type of categorical variable where the values fall into a hierarchical order (i.e. low/medium/high, disagree/neutral/agree)
- The term ‘categorical’ is often used without specifying ‘ordinal’, but it’s good to understand the technical distinction





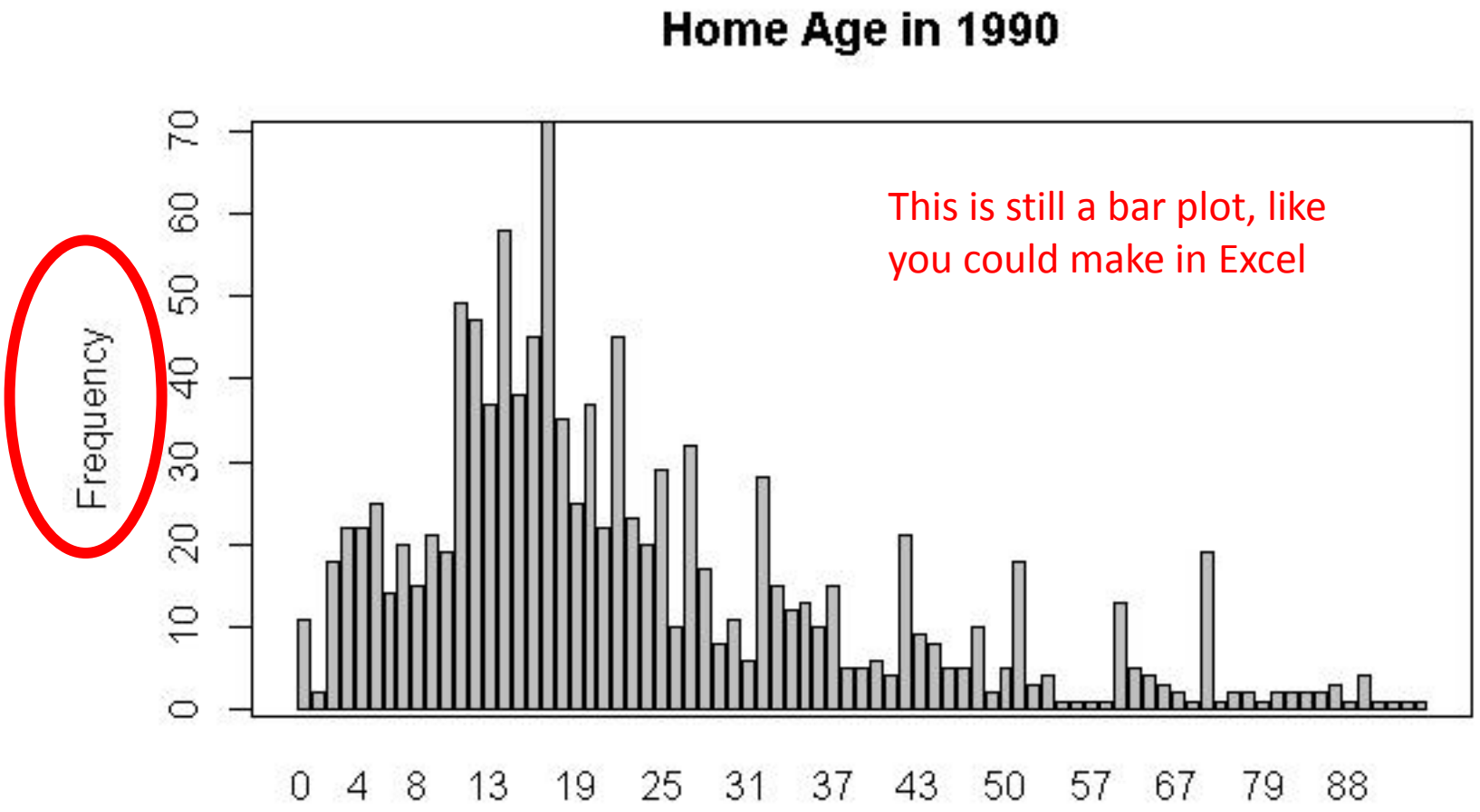
What's on the Y-axis

- Can be expressed as counts or frequencies
- Can also be expressed as percentages, which indicate the probability of observing each category within the dataset
- Statistically, this is often referred to as 'density'



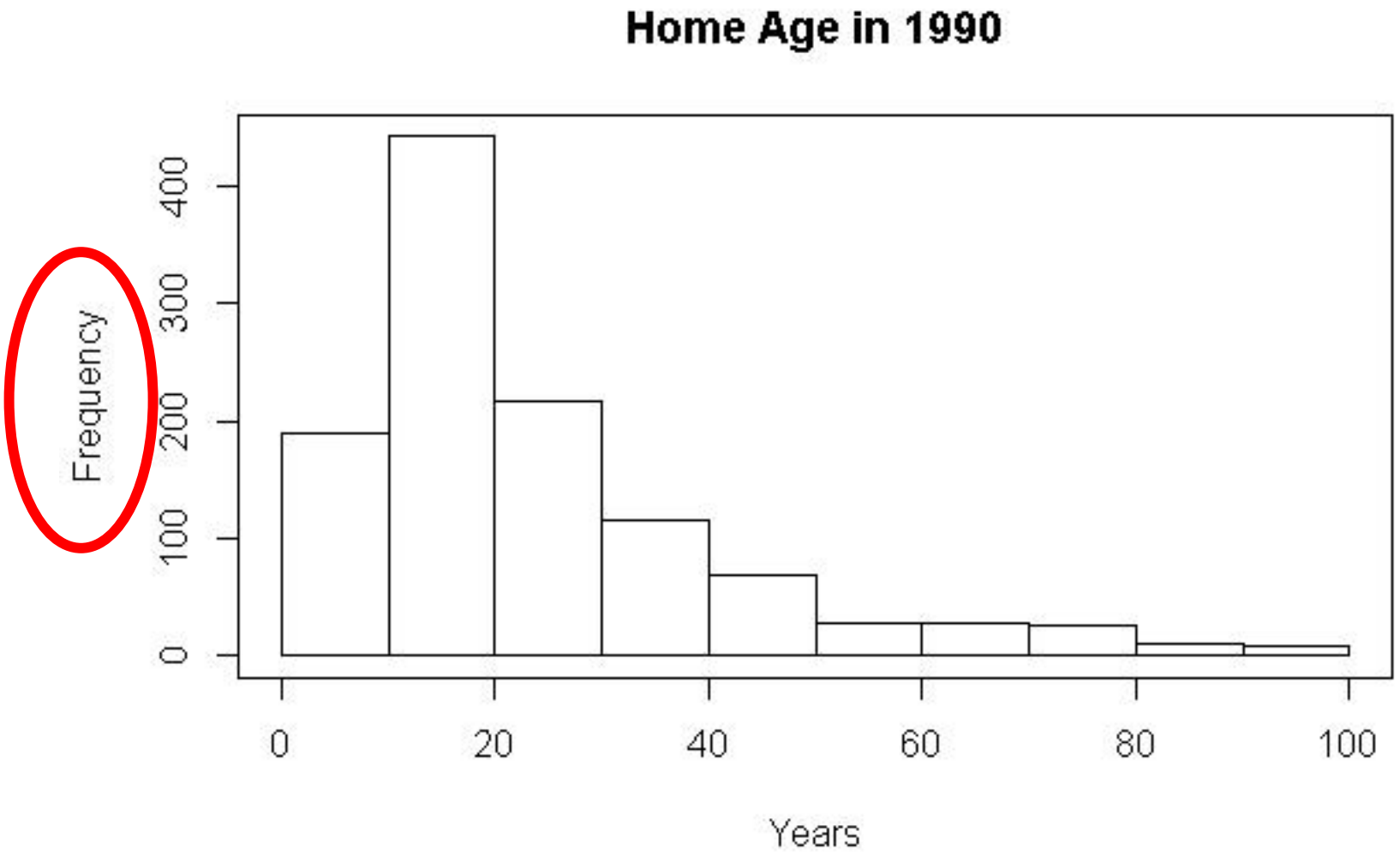
Continuous Variables

- Can take ANY value within their possible range (i.e. weight, radon concentration)
- Variables that can take an infinite number of integer values are often called ‘discrete’



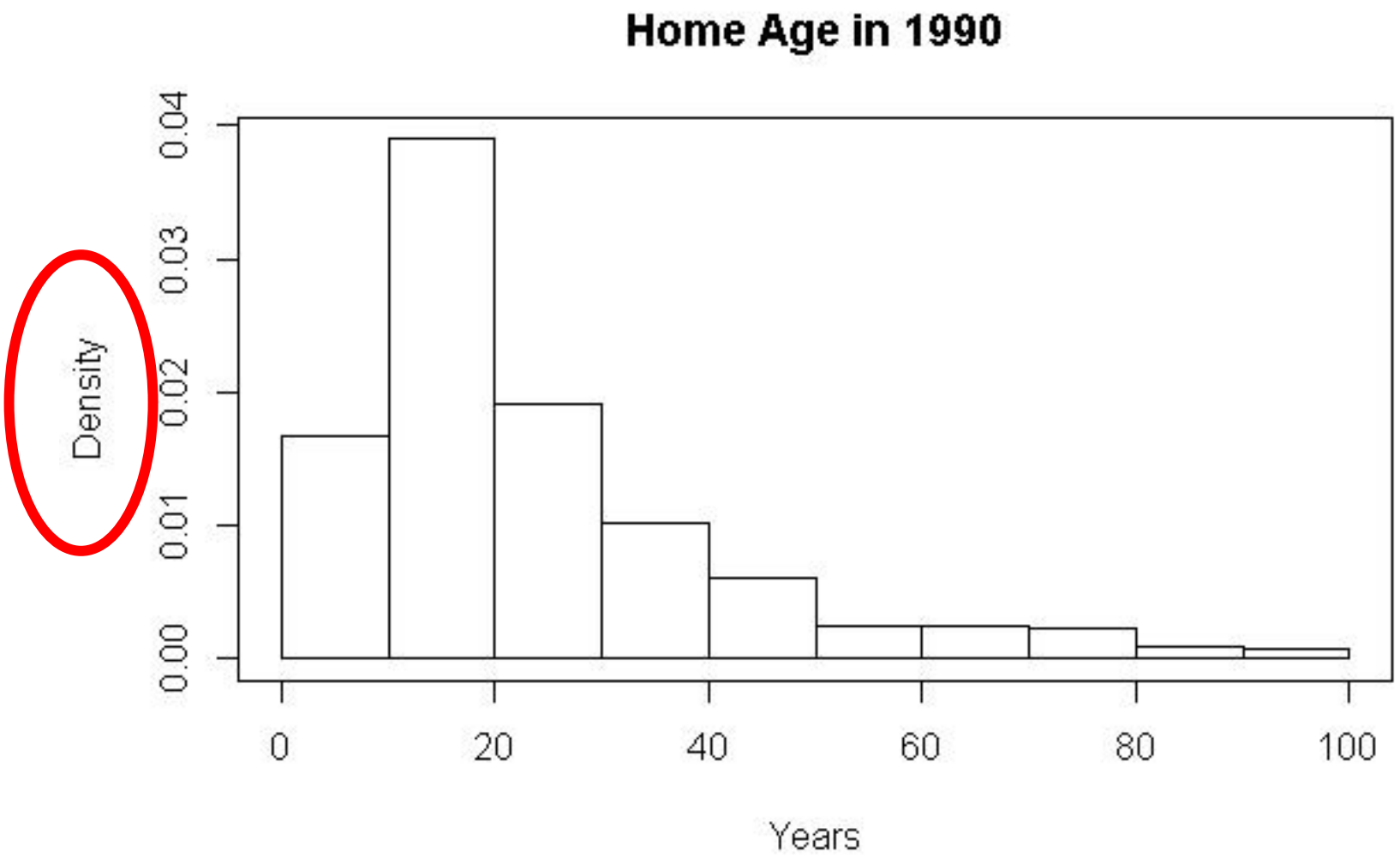
Frequency Histograms

- The standard way to visualize the values of a continuous variable
- All statistic programs will have a way to generate histograms



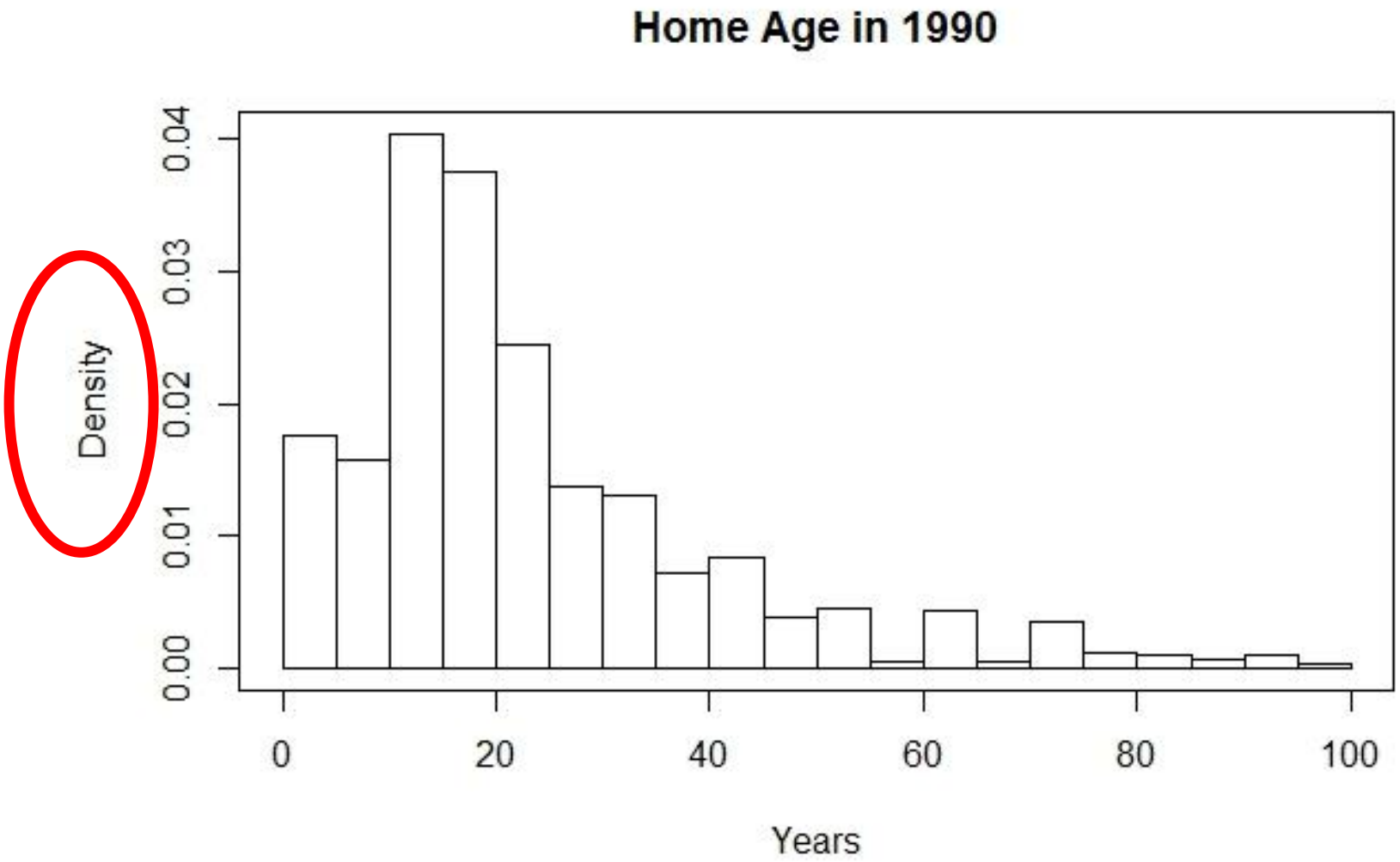
Density Histograms

- Shows us the probability of observing values within group for the entire dataset



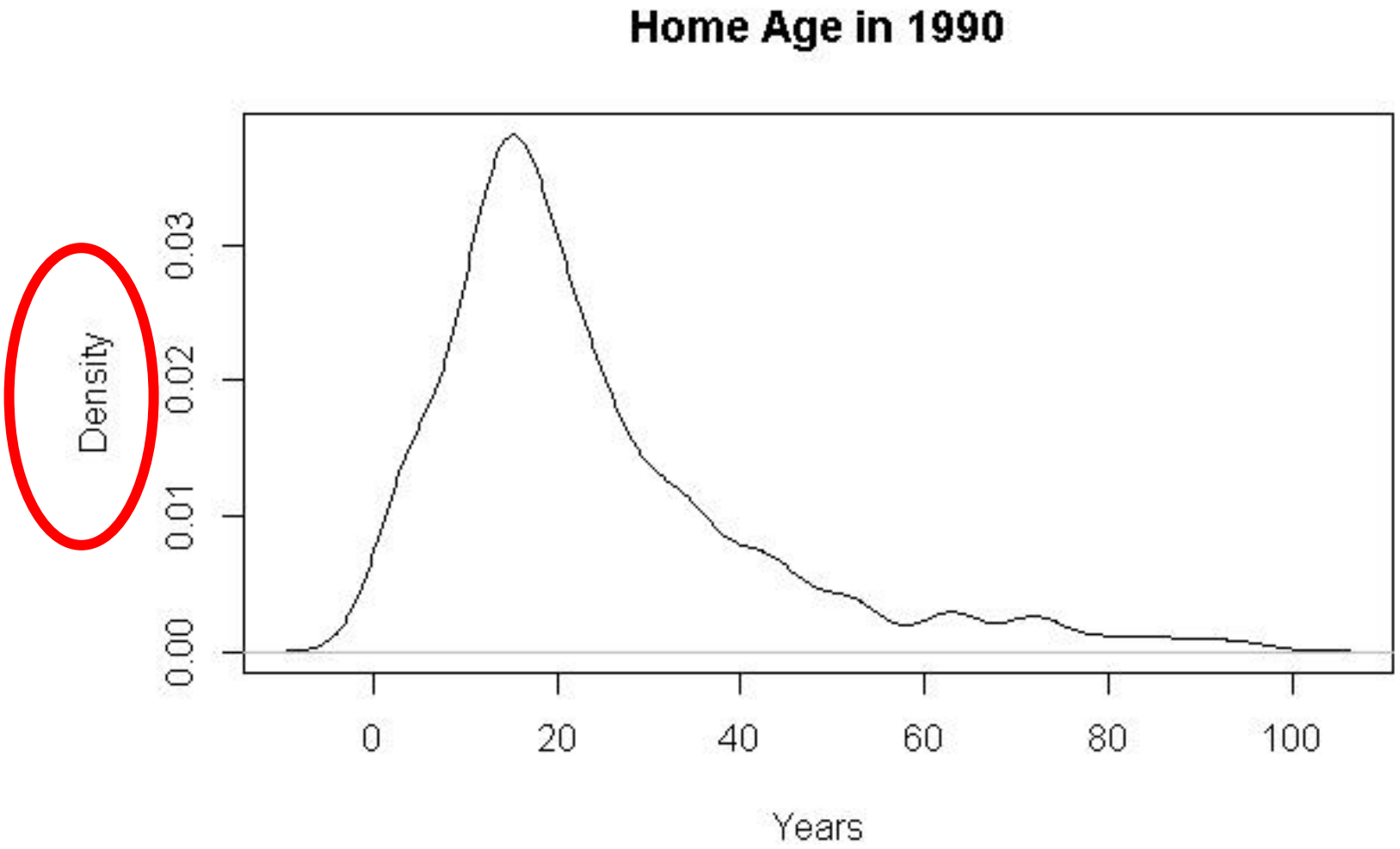
Density Histograms

- You can control the width of the ranges over which values are aggregated to get a clearer picture



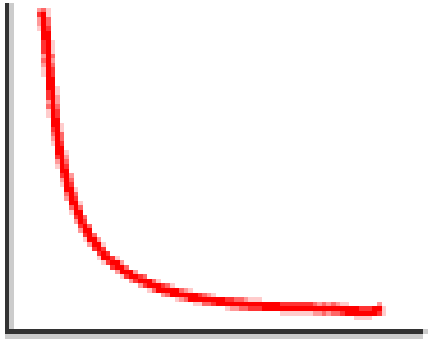
Density Functions

- Or you can fit a smooth line to the data to get a density function that will give you an exact estimate of the probability of observing any value in the range

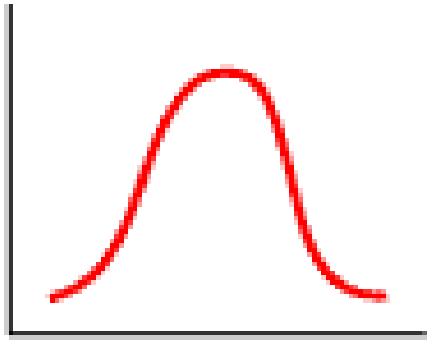


Probability Distributions

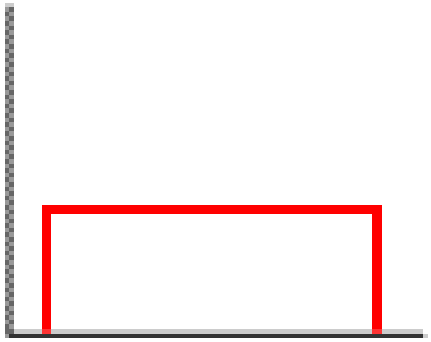
- Most continuous data approximate the shape of a STANDARD probability distribution, and this is why we can do statistics
- We will focus on parametric methods that assume our data follow some type of normal distribution



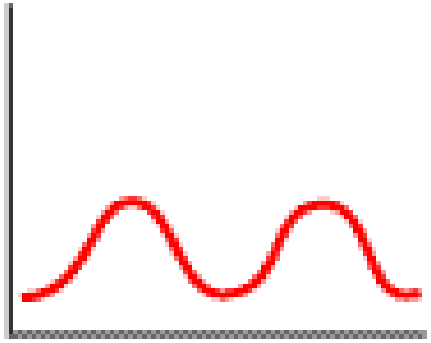
J-shaped



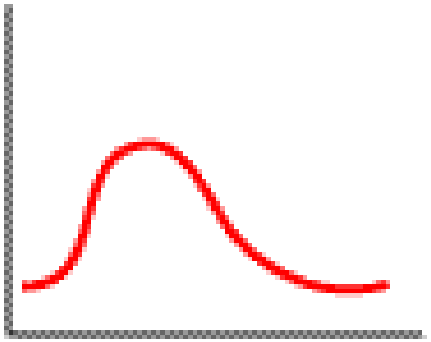
Normal



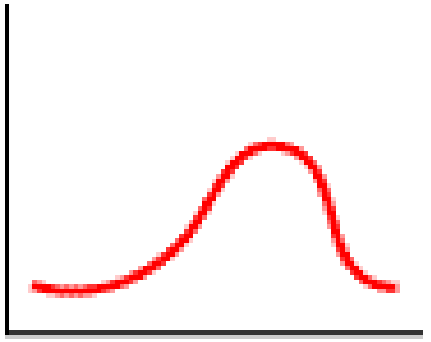
Rectangular



Bimodal



Positive (right) skew



Negative (left) skew

Logarithms Review

- What value do we have to put in the box to make the number in brackets?

$$2^{\square}(8) = ?$$

$$2^{\square}(16) = ?$$

$$2^{\square}(2) = ?$$

$$3^{\square}(27) = ?$$

$$3^{\square}(16) = ?$$

$$2^{\square}\left(\frac{1}{2}\right) = ?$$

$$10^{\square}(10000) = ?$$

What is a Logarithm?

- The logarithm of X is the number of times you have to multiply the base of the logarithm by itself to get X
- We will always use base *e*, which is the base of the NATURAL LOGARITHM, or approximately 2.71828

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} = 1 + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \cdots$$

- The mathematical constant *e* is used to describe exponential functions
- When I say LOG I always mean the natural logarithm, and you should too

Next Week

- Normal and lognormal distributions
- Summary statistics
- Skewness and kurtosis
- Examples in occupational and environmental health
- Censored values and what to do with them