Note: you will get a R code cheat sheet that includes the functions for the exam.

Imagine you are a consultant for a record company, and they have no idea what statistics are, but know that they need them. Given the following data, answer their questions with *R* output or in jargon free lingo. You can assume they know what a mean score consists of, but not necessarily things like variance or standard deviation.

We’ve recently collected data from our artists about their record sales, and weeks on the charts (see table below). We wanted some pretty pictures and descriptions of the data to be able to put into our presentation for the CEO in a couple weeks. Record sales are in the thousands.

|  |  |  |
| --- | --- | --- |
| Type of Artist | Record Sales | Chart Weeks |
| Pop | 105 | 5 |
| Country | 500 | 10 |
| Pop | 75 | 3 |
| Pop | 203 | 5 |
| Pop | 100 | 5 |
| Country | 109 | 5 |
| Country | 650 | 15 |
| Country | 400 | 4 |
| Pop | 65 | 3 |

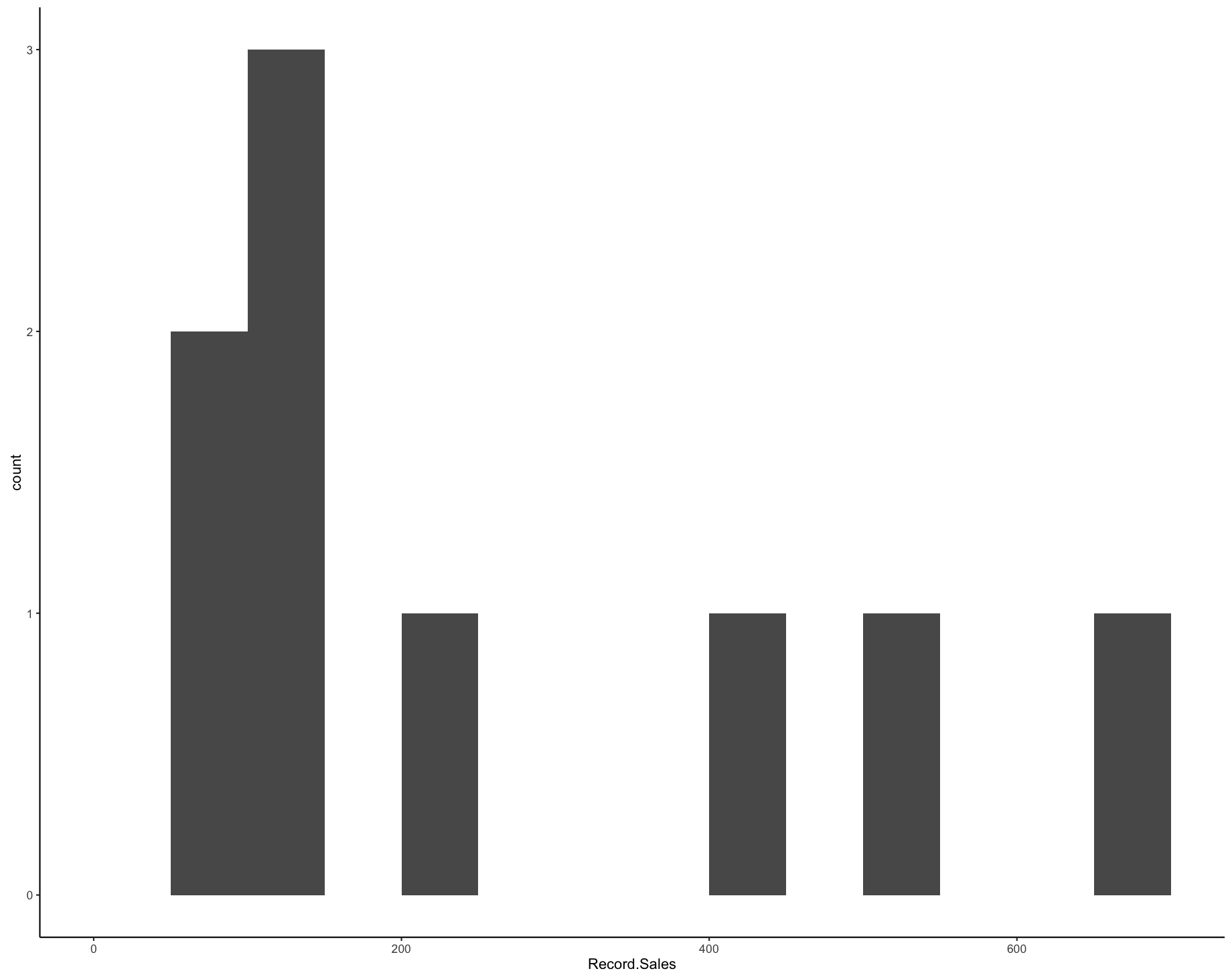
For the first column only:

* What is the variable? **Type of artist**
* What are the levels? **Pop and country**

For the rest of the columns:

* What type of variable is record sales using the NOIR system? **Ratio**
* How might the company operationally define chart weeks? **the weeks it was in the top 40.**

Create a histogram of record sales.



* Do there appear to be outliers? **Yes on the right**
* What shape is the histogram? **Unimodal**
* Does the distribution have any skew? **Positively skewed**

Create a frequency table of chart weeks.

**3 4 5 10 15**

**2 1 4 1 1**

What are the mean, median, mode(s), and interquartile range for record sales and chart weeks?

**Record sales**

**Mean and median with summary**

**> summary(reviewdata$Record.Sales)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**65.0 100.0 109.0 245.2 400.0 650.0**

**IQR**

**> IQR(reviewdata$Record.Sales, na.rm = T)**

**[1] 300**

**Mode(s)**

**> temp <- table(as.vector(reviewdata$Record.Sales))**

**> names(temp)[temp == max(temp)]**

**[1] "65" "75" "100" "105" "109" "203" "400" "500" "650"**

**Chart weeks**

**> summary(reviewdata$Chart.Weeks)**

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

**3.000 4.000 5.000 6.111 5.000 15.000**

**> IQR(reviewdata$Chart.Weeks, na.rm = T)**

**[1] 1**

**> temp <- table(as.vector(reviewdata$Chart.Weeks))**

**> names(temp)[temp == max(temp)]**

**[1] "5"**

What is the *biased* standard deviation and variance estimate for chart weeks?

**> pop.var(reviewdata$Chart.Weeks)**

**[1] 13.65432**

**> pop.sd(reviewdata$Chart.Weeks)**

**[1] 3.695175**

How does that differ than the *unbiased* estimates (you can describe rather than give numbers)?

**The difference is the denominator (biased N on the bottom and is for populations, while unbiased has N - 1 on the bottom and is for samples).**

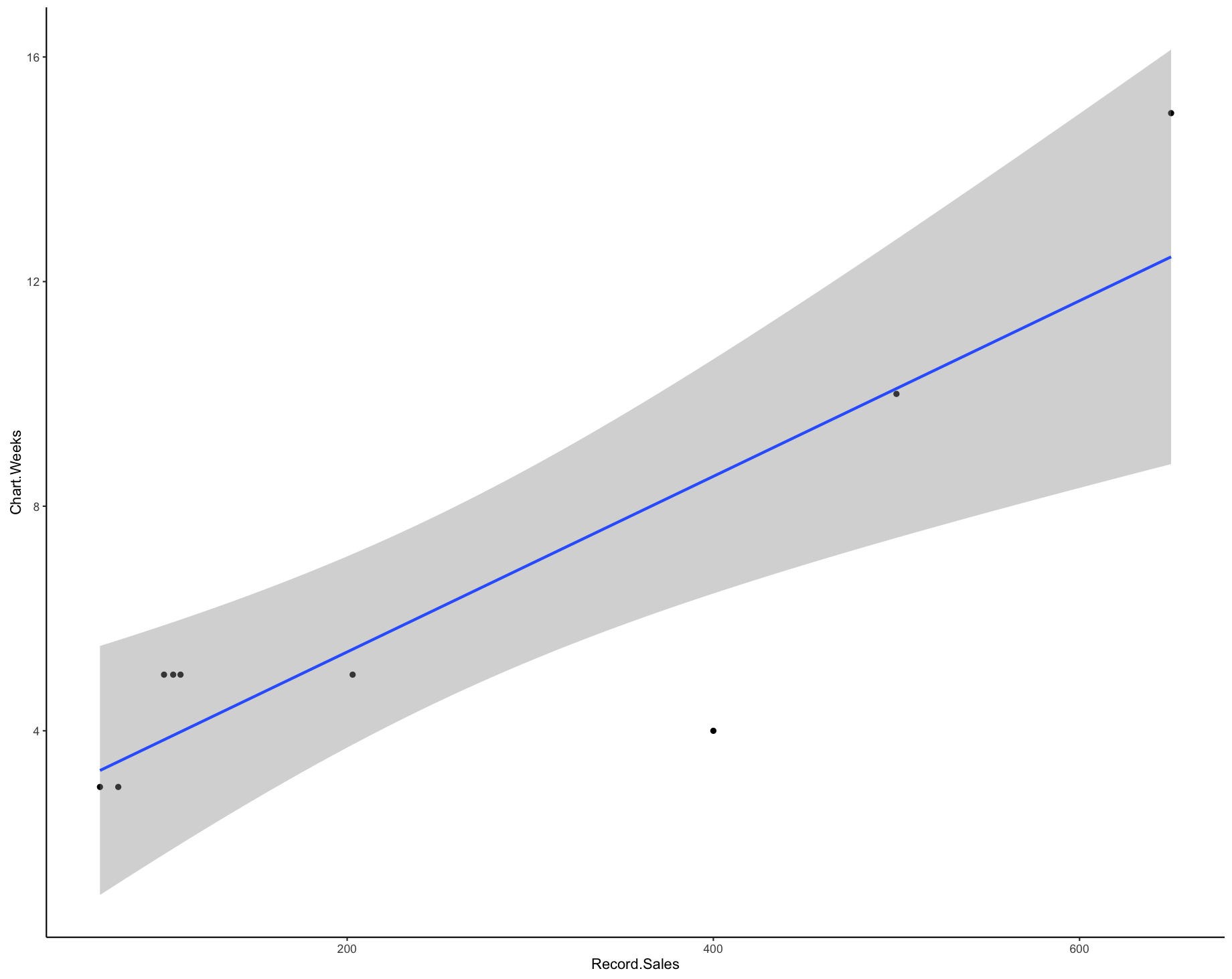
**> var(reviewdata$Chart.Weeks, na.rm = T)**

**[1] 15.36111**

**> sd(reviewdata$Chart.Weeks, na.rm = T)**

**[1] 3.919325**

Create a graph of record sales and chart weeks with a line of best fit.



Create a graph for the different types of artists’ average record sales. Who seems to sell more records?

