Week 2: Data Summary and Visualization

1. Categorical variables I

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

The categorical variables identify a category for each case. They have a limited number of different values, called **levels**.

Examples:
If our observational units are 100 M&M's. Color of a M&M is a categorical variable.

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

Outline

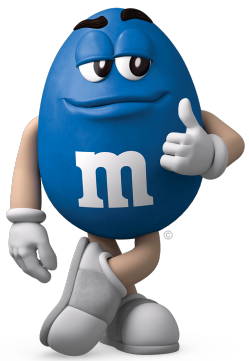
1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

Distribution is about "pattern of variation in a variable"
The key of distribution is to focus on the variation across the entire data set.



individual level
color of the MM is blue

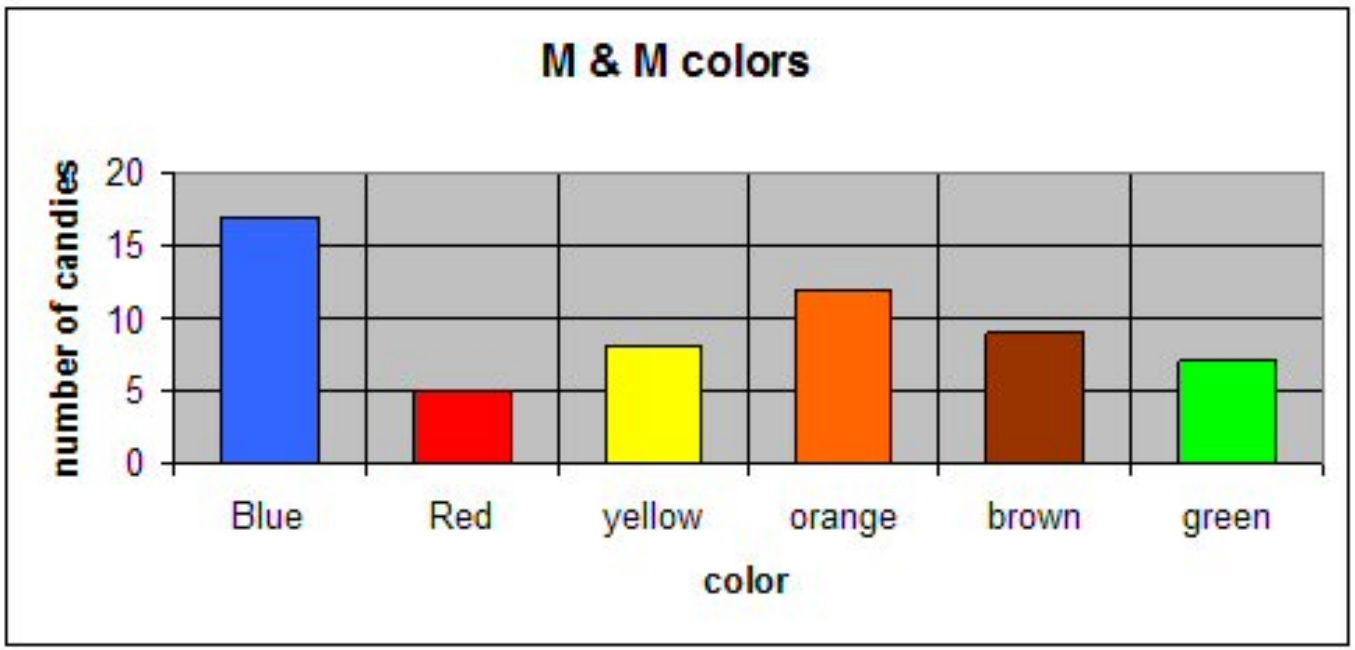
macro level
colors of all MM's

In this class, we focus on distribution of **categorical variable**.

For a categorical variable, the distribution

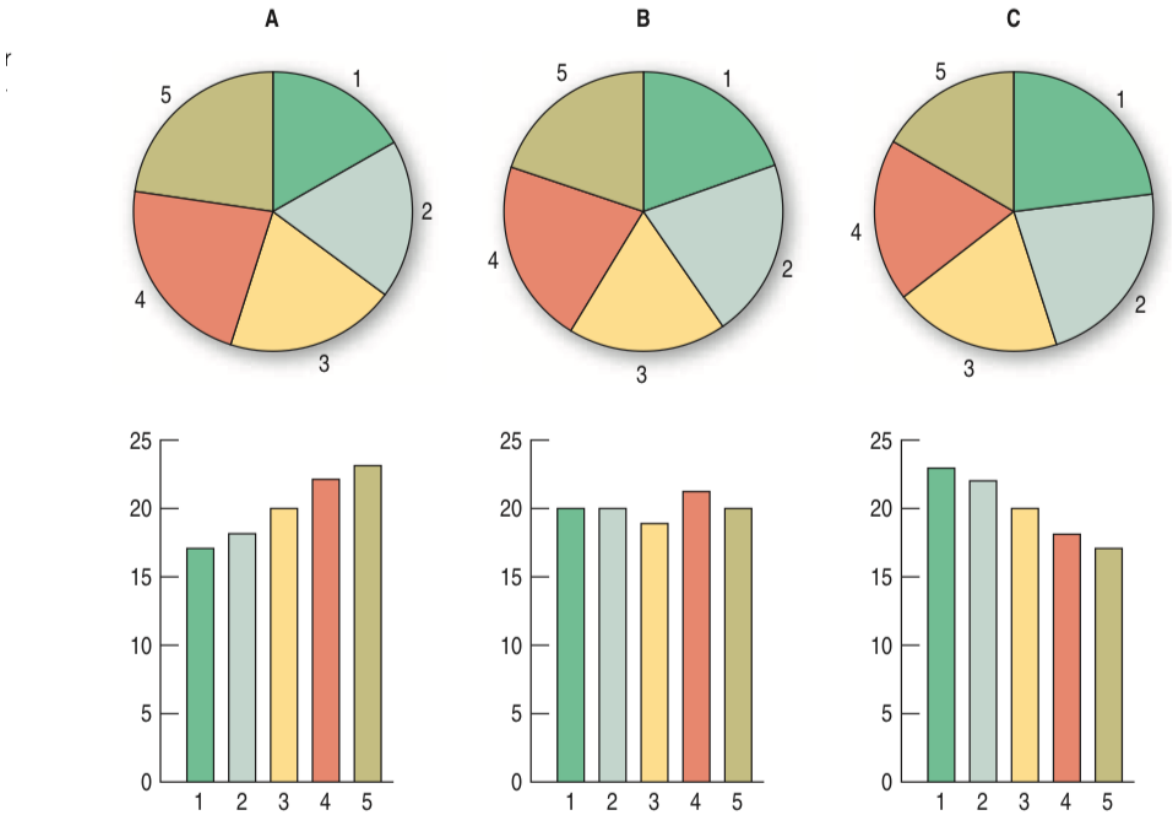
- names the possible levels in a categorical variable
- tells how many times each occurs

We can visualize the distribution by a bar plot or a pie char. In this course, we focus on bar plot.



Pie chart is not good for visualizing the relative frequencies

Three pie charts that look pretty much alike along with bar charts of the same data. The bar charts show three distinctly different patterns, but it is almost impossible to see those in the pie charts.



Suppose we have a data set that consists of almost 54,000 diamonds that contains the prices and their attributes (e.g cut, color, clarity).

Poll question

Which command would help you take a glimpse of the data?

1. ncol(diamonds)
2. nrow(diamonds)
3. glimpse(diamonds)

The data table is given as follow

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>
0.23	Ideal	E	SI2	61.5	55	326
0.21	Premium	E	SI1	59.8	61	326
0.23	Good	E	VS1	56.9	65	327
0.29	Premium	I	VS2	62.4	58	334
0.31	Good	J	SI2	63.3	58	335
0.24	Very Good	J	VVS2	62.8	57	336

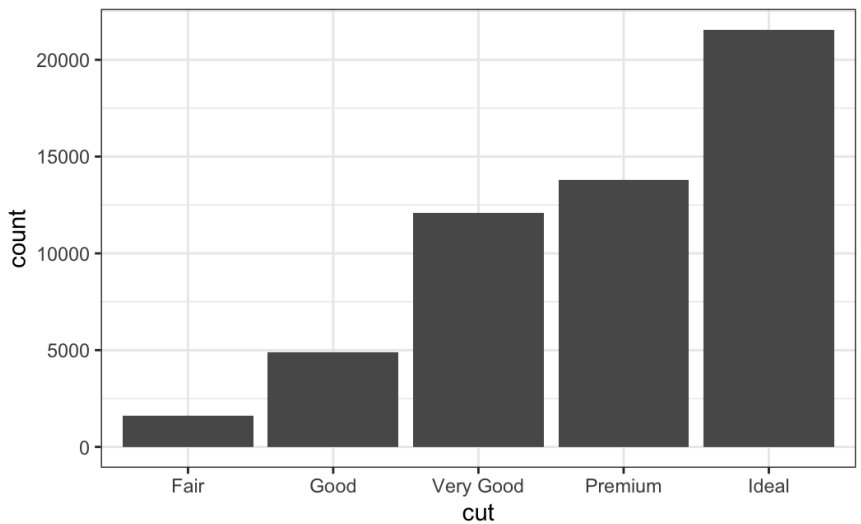
6 rows | 1-7 of 10 columns

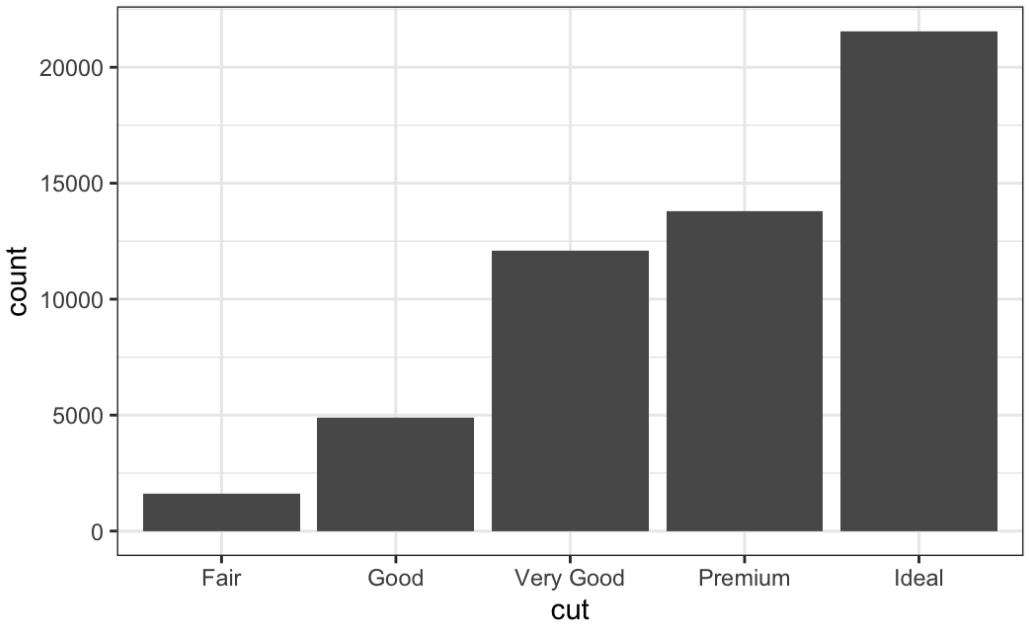
When a variable is categorical, you can visualize the distribution with a bar plot. There are lots of ways to make bar plots in R. Throughout this course, we will use the package ‘ggplot2‘ to make our visualizations.

For example, the code below plots a bar plot of the ‘cut‘ variable.

Rcode

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
```

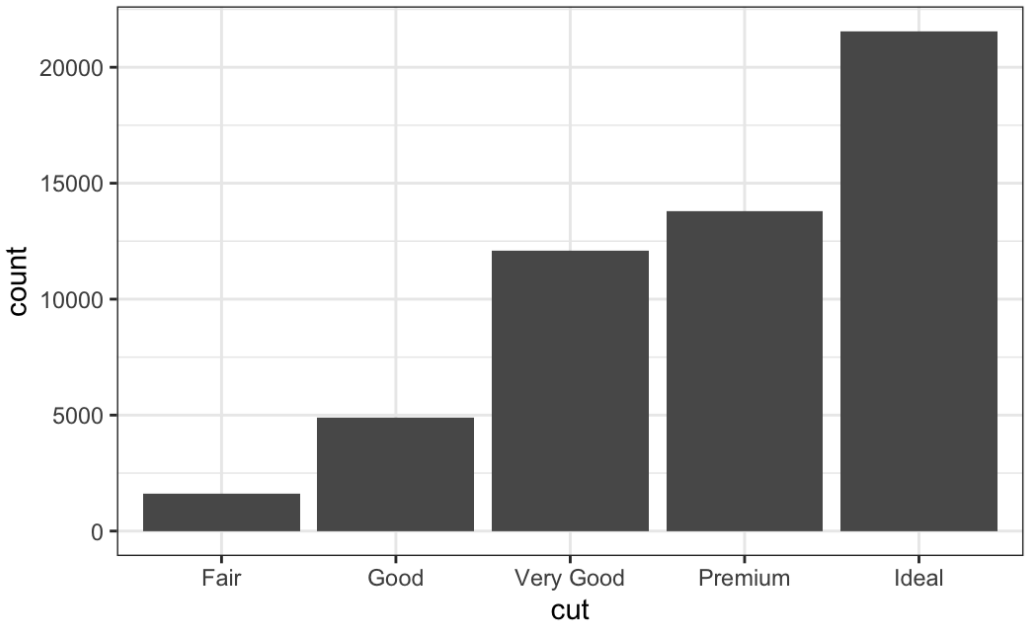




Poll question

What does the y axis represent in the bar plot?

1. Count of diamond in each category
2. Frequencies of diamond in each category
3. The total number of diamonds



Poll question

How many diamonds in the dataset had a ‘Good‘ cut?

1. 2000
2. 5000
3. 7000

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

The bar plot displays the counts – how many times each value appears in the data – on the y-axis. You can also organize this information in a **frequency table**. Note the term frequency refers to the number of times each value occurs in the respective data set.

cut <ord>	**n** <int>
Fair	1610
Good	4906
Very Good	12082
Premium	13791
Ideal	21551

Using the frequency table, one can calculate a notion of proportion.

For example, in the diamond dataset, if we know there are 53,940 units in the sample, and we know 21,551 of them have ‘ideal‘ cut. Calculate the proportion of diamonds that have ‘Ideal‘ cut.

$$\frac{21,551}{53,940} = 0.3995 = 40\%$$

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

If we are interested in knowing the relationship between categorical variables, we can look at the contingency table.

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

The entries in the table count how many observational units are in each combination of levels of the cut and color variables. If you add up all the numbers in the table, you will get the total number of observational units in the sample.

Calculate the proportion of diamonds that have 'cut' *Fair* **and** 'color' *E*

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Calculate the proportion of diamonds fall in ‘cut‘ *Fair* (aggregating across all values of ‘color‘)

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Among those cases where the diamonds have ‘cut‘ *Fair* , calculate the proportion of diamonds that have ‘color‘ *E*.

color <ord>	Fair <int>	Good <int>	Very Good <int>	Premium <int>	Ideal <int>
D	163	662	1513	1603	2834
E	224	933	2400	2337	3903
F	312	909	2164	2331	3826
G	314	871	2299	2924	4884
H	303	702	1824	2360	3115
I	175	522	1204	1428	2093
J	119	307	678	808	896

Outline

1. Questions? 15 min

2. Today: Categorical variables

3. Main ideas
    1. Distributions
    2. Proportional reasoning
    3. Contingency table

4. Summary

1. Distributions
2. Proportional reasoning
3. Contingency table