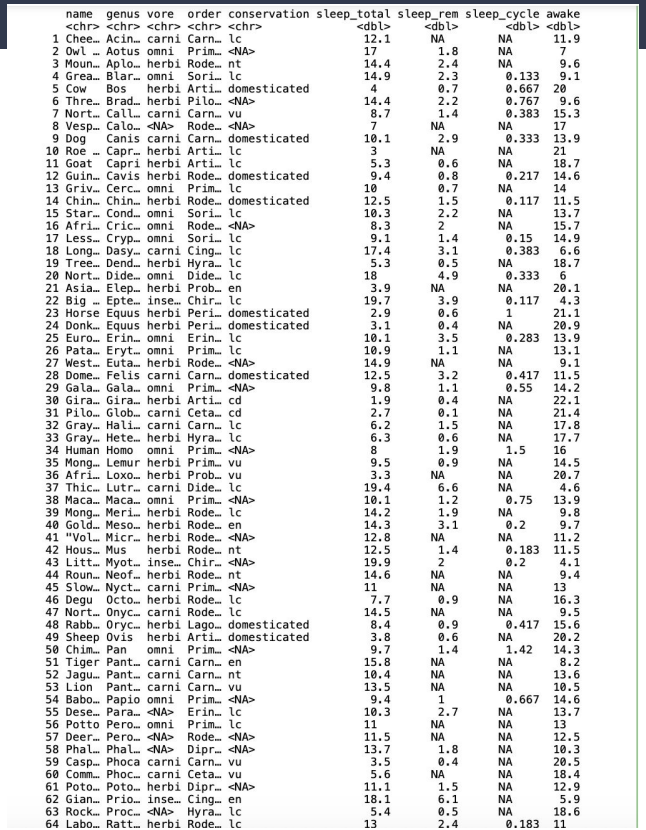
Exploratory Data Analysis

BSDS 100, Spring 2021

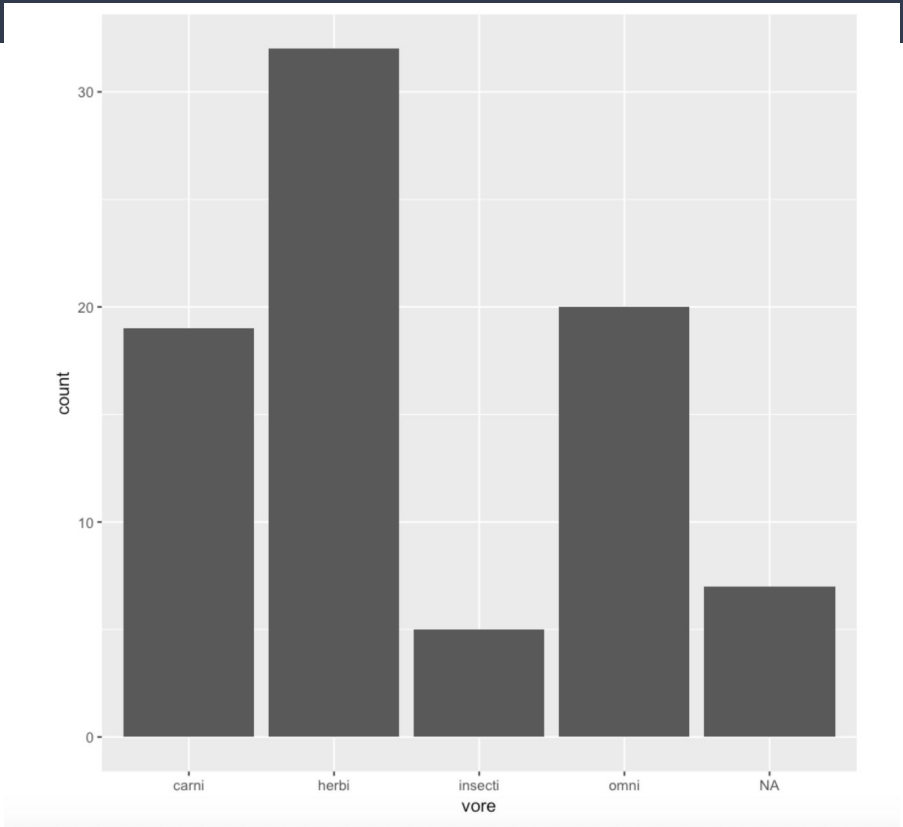
Michael Ruddy

What is Exploratory Data Analysis?



	name	genus	vore	order	conservation	sleep_total	sleep_rem	sleep_cycle	awake
	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	Chee…	Acin…	carni	Carn…	lc	12.1	NA	NA	11.9
2	Owl …	Aotus	omni	Prim…	<NA>	17	1.8	NA	7
3	Moun…	Aplo…	herbi	Rode…	nt	14.4	2.4	NA	9.6
4	Grea…	Blar…	omni	Sori…	lc	14.9	2.3	0.133	9.1
5	Cow	Bos	herbi	Arti…	domesticated	4	0.7	0.667	20
6	Thre…	Brad…	herbi	Pilo…	<NA>	14.4	2.2	0.767	9.6
7	Nort…	Call…	carni	Carn…	vu	8.7	1.4	0.383	15.3
8	Vesp…	Calo…	<NA>	Rode…	<NA>	7	NA	NA	17
9	Dog	Canis	carni	Carn…	domesticated	10.1	2.9	0.333	13.9
10	Roe …	Capr…	herbi	Arti…	lc	3	NA	NA	21
11	Goat	Capri	herbi	Arti…	lc	5.3	0.6	NA	18.7
12	Guin…	Cavis	herbi	Rode…	domesticated	9.4	0.8	0.217	14.6
13	Griv…	Cerc…	omni	Prim…	lc	10	0.7	NA	14
14	Chin…	Chin…	herbi	Rode…	domesticated	12.5	1.5	0.117	11.5
15	Star…	Cond…	omni	Sori…	lc	10.3	2.2	NA	13.7
16	Afri…	Cric…	omni	Rode…	<NA>	8.3	2	NA	15.7
17	Less…	Cryp…	omni	Sori…	lc	9.1	1.4	0.15	14.9
18	Long…	Dasy…	carni	Cing…	lc	17.4	3.1	0.383	6.6
19	Tree…	Dend…	herbi	Hyra…	lc	5.3	0.5	NA	18.7
20	Nort…	Dide…	omni	Dide…	lc	18	4.9	0.333	6
21	Asia…	Elep…	herbi	Prob…	en	3.9	NA	NA	20.1
22	Big …	Epte…	inse…	Chir…	lc	19.7	3.9	0.117	4.3
23	Horse	Equus	herbi	Peri…	domesticated	2.9	0.6	1	21.1
24	Donk…	Equus	herbi	Peri…	domesticated	3.1	0.4	NA	20.9
25	Euro…	Erin…	omni	Erin…	lc	10.1	3.5	0.283	13.9
26	Pata…	Eryt…	omni	Prim…	lc	10.9	1.1	NA	13.1
27	West…	Euta…	herbi	Rode…	<NA>	14.9	NA	NA	9.1
28	Dome…	Felis	carni	Carn…	domesticated	12.5	3.2	0.417	11.5
29	Gala…	Gala…	omni	Prim…	<NA>	9.8	1.1	0.55	14.2
30	Gira…	Gira…	herbi	Arti…	cd	1.9	0.4	NA	22.1
31	Pilo…	Glob…	carni	Ceta…	cd	2.7	0.1	NA	21.4
32	Gray…	Hali…	carni	Carn…	lc	6.2	1.5	NA	17.8
33	Gray…	Hete…	herbi	Hyra…	lc	6.3	0.6	NA	17.7
34	Human	Homo	omni	Prim…	<NA>	8	1.9	1.5	16
35	Mong…	Lemur	herbi	Prim…	vu	9.5	0.9	NA	14.5
36	Afri…	Loxo…	herbi	Prob…	vu	3.3	NA	NA	20.7
37	Thic…	Lutr…	carni	Dide…	lc	19.4	6.6	NA	4.6
38	Maca…	Maca…	omni	Prim…	<NA>	10.1	1.2	0.75	13.9
39	Mong…	Meri…	herbi	Rode…	lc	14.2	1.9	NA	9.8
40	Gold…	Meso…	herbi	Rode…	en	14.3	3.1	0.2	9.7
41	"Vol…	Micr…	herbi	Rode…	<NA>	12.8	NA	NA	11.2
42	Hous…	Mus	herbi	Rode…	nt	12.5	1.4	0.183	11.5
43	Litt…	Myot…	inse…	Chir…	<NA>	19.9	2	0.2	4.1
44	Roun…	Neof…	herbi	Rode…	nt	14.6	NA	NA	9.4
45	Slow…	Nyct…	carni	Prim…	<NA>	11	NA	NA	13
46	Degu	Octo…	herbi	Rode…	lc	7.7	0.9	NA	16.3
47	Nort…	Onyc…	carni	Rode…	lc	14.5	NA	NA	9.5
48	Rabb…	Oryc…	herbi	Lago…	domesticated	8.4	0.9	0.417	15.6
49	Sheep	Ovis	herbi	Arti…	domesticated	3.8	0.6	NA	20.2
50	Chim…	Pan	omni	Prim…	<NA>	9.7	1.4	1.42	14.3
51	Tiger	Pant…	carni	Carn…	en	15.8	NA	NA	8.2
52	Jagu…	Pant…	carni	Carn…	nt	10.4	NA	NA	13.6
53	Lion	Pant…	carni	Carn…	vu	13.5	NA	NA	10.5
54	Babo…	Papio	omni	Prim…	<NA>	9.4	1	0.667	14.6
55	Dese…	Para…	<NA>	Erin…	lc	10.3	2.7	NA	13.7
56	Potto	Pero…	omni	Prim…	lc	11	NA	NA	13
57	Deer…	Pero…	<NA>	Rode…	<NA>	11.5	NA	NA	12.5
58	Phal…	Phal…	<NA>	Dipr…	<NA>	13.7	1.8	NA	10.3
59	Casp…	Phoca	carni	Carn…	vu	3.5	0.4	NA	20.5
60	Comm…	Phoc…	carni	Ceta…	vu	5.6	NA	NA	18.4
61	Poto…	Poto…	herbi	Dipr…	<NA>	11.1	1.5	NA	12.9
62	Gian…	Prio…	inse…	Cing…	en	18.1	6.1	NA	5.9
63	Rock…	Proc…	<NA>	Hyra…	lc	5.4	0.5	NA	18.6
64	Labo…	Ratt…	herbi	Rode…	lc	13	2.4	0.183	11

What is Exploratory Data Analysis?



What is Exploratory Data Analysis?

- **Informal** exploration of your data
- Summarize and visualize properties of your dataset
- Iterative Procedure:
    - 1. Generate questions, hypotheses
    - 2. Visualize, transform, model your data
    - 3. Refine your questions, repeat

Terms

- **Variable**: a quantity, quality, or property that you can measure
- **Value**: the state of a variable when you measure it
- **Observation**: set of a measurements under similar conditions (time, object, etc.); several values for different variables
- **Tabular Data**: Set of values, each associated with a variable and an observation. (Hopefully its *tidy*).

Terms

Tabular Data

Variables → name, genus, vore, order, conservation, sleep_total, sleep_rem, sleep_cycle, awake, brainwt, bodywt

name	genus	vore	order	conservation	sleep_total	sleep_rem	sleep_cycle	awake	brainwt	bodywt
Cheetah	Acinonyx	carni	Carnivora	lc	12.1	NA	NA	11.9	NA	50.000
Owl monkey	Aotus	omni	Primates	NA	17.0	1.8	NA	7.0	0.01550	0.480
Mountain beaver	Aplodontia	herbi	Rodentia	nt	14.4	2.4	NA	9.6	NA	1.350
Greater short-tailed shrew	Blarina	omni	Soricomorpha	lc	14.9	2.3	0.1333333	9.1	0.00029	0.019
Cow	Bos	herbi	Artiodactyla	domesticated	4.0	0.7	0.6666667	20.0	0.42300	600.000
Three-toed sloth	Bradypus	herbi	Pilosa	NA	14.4	2.2	0.7666667	9.6	NA	3.850

Observation → Cow, Bos, herbi, Artiodactyla, domesticated, 4.0, 0.7, 0.6666667, 20.0, 0.42300, 600.000

Values

Types of Variables (non-exhaustive)

- **Quantitative**: variables representing numerical values you can perform arithmetic operations with.
  - *Discrete*: integer numbers (nothing “in-between”)
  - *Continuous*: real numbers
- **Qualitative (categorical)**: variables representing non-numeric properties
  - *Nominal*: No rank or ordering
  - *Ordered*: Clear rank or ordering

Types of Variables (non-exhaustive)

Nominal Discrete Ordered Continuous

name	year	month	day	hour	lat	long	status	category	wind	pressure	ts_diameter	hu_diameter
Amy	1975	6	27	0	27.5	-79.0	tropical depression	-1	25	1013	NA	NA
Amy	1975	6	27	6	28.5	-79.0	tropical depression	-1	25	1013	NA	NA
Amy	1975	6	27	12	29.5	-79.0	tropical depression	-1	25	1013	NA	NA
Amy	1975	6	27	18	30.5	-79.0	tropical depression	-1	25	1013	NA	NA
Amy	1975	6	28	0	31.5	-78.8	tropical depression	-1	25	1012	NA	NA
Amy	1975	6	28	6	32.4	-78.7	tropical depression	-1	25	1012	NA	NA
Amy	1975	6	28	12	33.3	-78.0	tropical depression	-1	25	1011	NA	NA
Amy	1975	6	28	18	34.0	-77.0	tropical depression	-1	30	1006	NA	NA
Amy	1975	6	29	0	34.4	-75.8	tropical storm	0	35	1004	NA	NA
Amy	1975	6	29	6	34.0	-74.8	tropical storm	0	40	1002	NA	NA
Amy	1975	6	29	12	33.8	-73.8	tropical storm	0	45	1000	NA	NA

Basic framework for questions

- What type of variation occurs within a variable?
  - Typical, unusual, or missing values
  - Numeric: mean, standard deviation, interquartile range, etc.
- What type of covariation occurs between the variables?
  - How are two or more variables related?
  - Correlation coefficient, covariance matrices, etc.

Variation with one variable: Categorical

- How many of each category? Proportion of total?
- Which values are typical? Uncommon? Missing?
    - Why? Is this expected?
- Summarize: count, proportion, percentage
- Visualize: bar graph

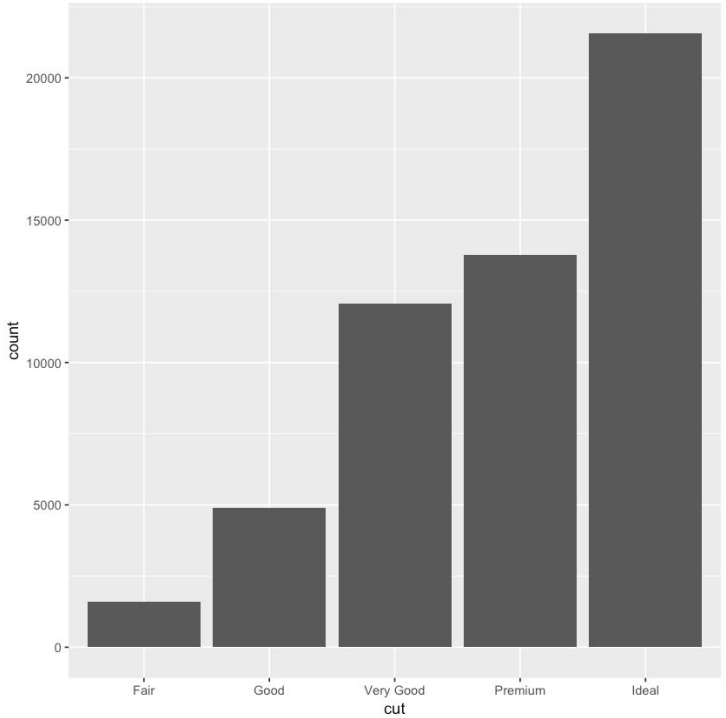
Variation with one variable: Categorical

In [31]:
```
diamonds %>%
  count(cut) %>%   # create a dataset with each value of "cut" variable and the count
  mutate(proportion = n / nrow(diamonds)) %>%   # add a column for proportion
  mutate(percent = proportion * 100)   # add a column for percent
```

cut	n	proportion	percent
Fair	1610	0.02984798	2.984798
Good	4906	0.09095291	9.095291
Very Good	12082	0.22398962	22.398962
Premium	13791	0.25567297	25.567297
Ideal	21551	0.39953652	39.953652

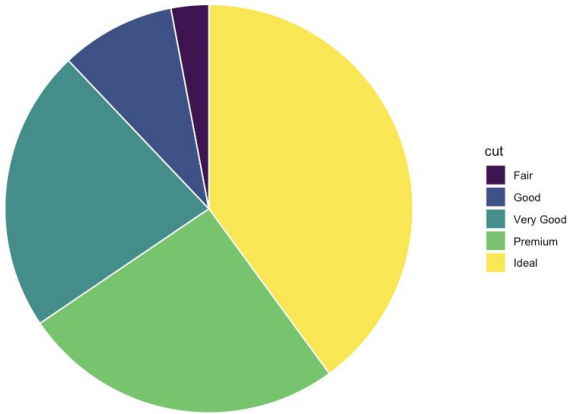
Variation with one variable: Categorical

```
In [33]: diamonds %>%
           ggplot(aes(x = cut)) +
           geom_bar()
```



Variation with one variable: Categorical

In [54]:
```
# create data with count of each cut
# create a bar graph with one bar, filled by cut (width 1, outline white to look nice)
# make it a polar chart
# remove axes and background
diamonds %>%
  count(cut) %>%  # create data with count of each cut
  ggplot(aes(x = "", y = n, fill = cut)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +
  theme(axis.line=element_blank(),axis.text.x=element_blank(),
          axis.text.y=element_blank(),axis.ticks=element_blank(),
          axis.title.x=element_blank(),
          axis.title.y=element_blank(),
          panel.background=element_blank(),
          panel.grid.major=element_blank(),
          panel.grid.minor=element_blank())
```



Variation with one variable: Numeric

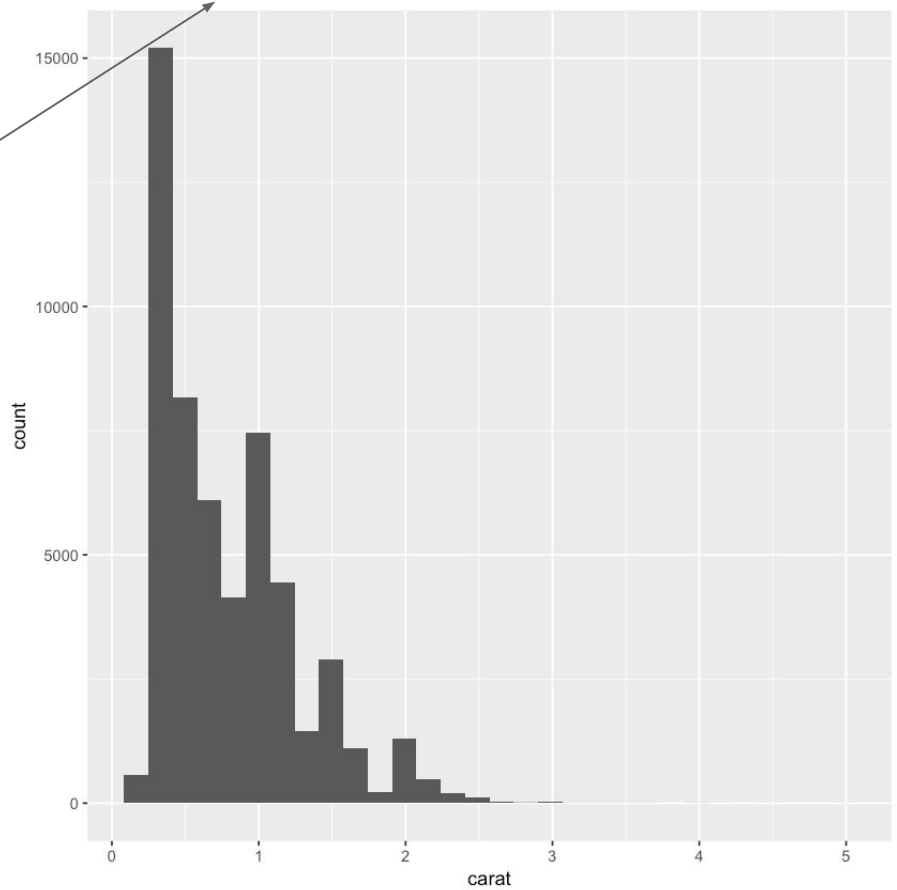
- Sometimes integer variables can be treated as categorical
- What is the distribution of values?
- Which values are typical? Uncommon? Missing?
  - Why? Is this expected?
- Summarize: mean, std. Dev, ...
- Visualize: histogram, box plot

Variation with one variable: Numeric

In [60]:
```
diamonds %>%
  summarize(mean = mean(carat), standard_dev = sd(carat)) %>%   # compute statistics
  mutate(name = "carat") %>%   # add the name column
  select(name, mean, standard_dev)   # rearrange order
```

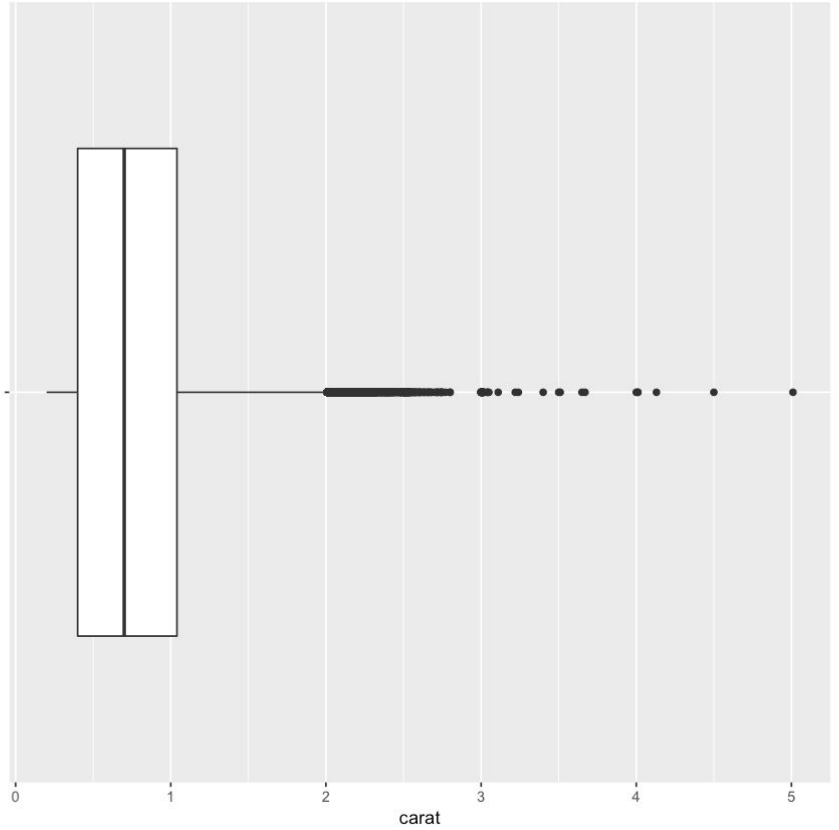
name	mean	standard_dev
carat	0.7979397	0.4740112

```
In [64]: diamonds %>%
           ggplot(aes(x = carat)) +
           geom_histogram(bins = 30)
```



Alter bins (or binwidth) for a finer or coarser look at the data.

In [77]:
```
# x controls the grouping of boxplots (can be empty or discrete/categorical)
# y controls the continuous variable
# flipping the box plot can make it look nicer
diamonds %>%
  ggplot(aes(x = "", y = carat)) +
  geom_boxplot() +
  theme(axis.text.y=element_blank(),
        axis.title.y=element_blank()) +
  coord_flip()
```



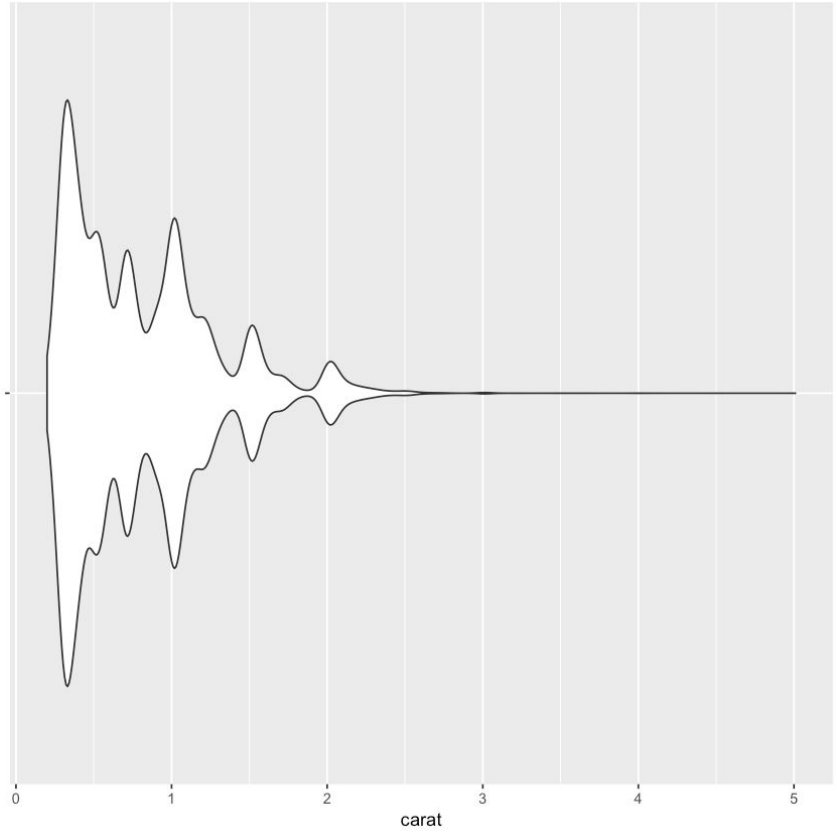
Computed variables

width
width of boxplot
ymin
lower whisker = smallest observation greater than or equal to lower hinge - 1.5 * IQR
lower
lower hinge, 25% quantile

middle
median, 50% quantile

upper
upper hinge, 75% quantile
ymax
upper whisker = largest observation less than or equal to upper hinge + 1.5 * IQR

In [83]:
```
diamonds %>%
  ggplot(aes(x = "", y = carat)) +
  geom_violin() +
  theme(axis.text.y=element_blank(),
        axis.title.y=element_blank()) +
  coord_flip()
```

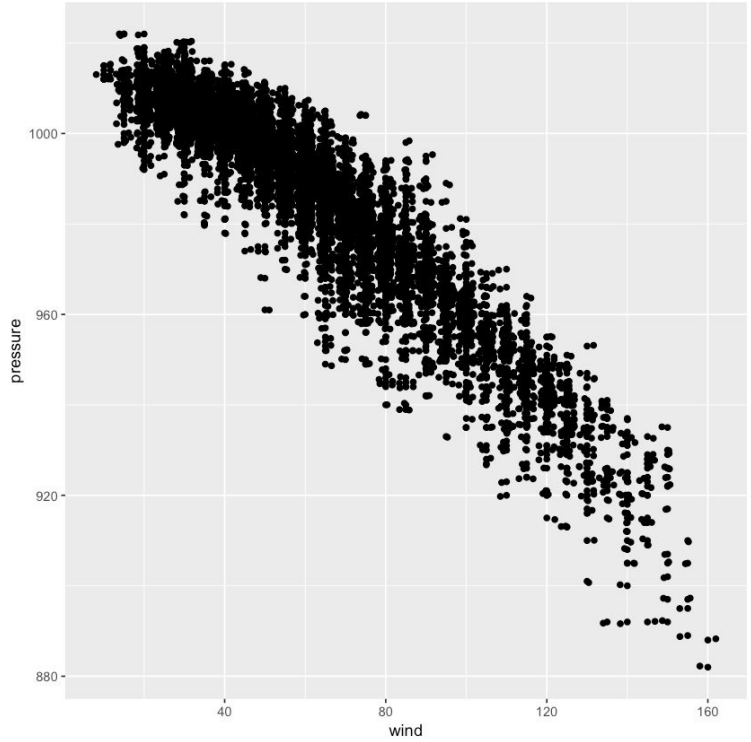


Covariation

- How are different variables related?
- Numeric vs. Numeric
  - Visualize: 2D scatter plot
  - Vs. Categorical: facets, scatter plot aesthetics
- Numeric vs. Categorical
  - Numeric variation tools by category
- Categorical vs. Categorical
  - Summarize: counts by category
  - Visualize: fancy bar graph, cross tables

Numeric vs. Numeric

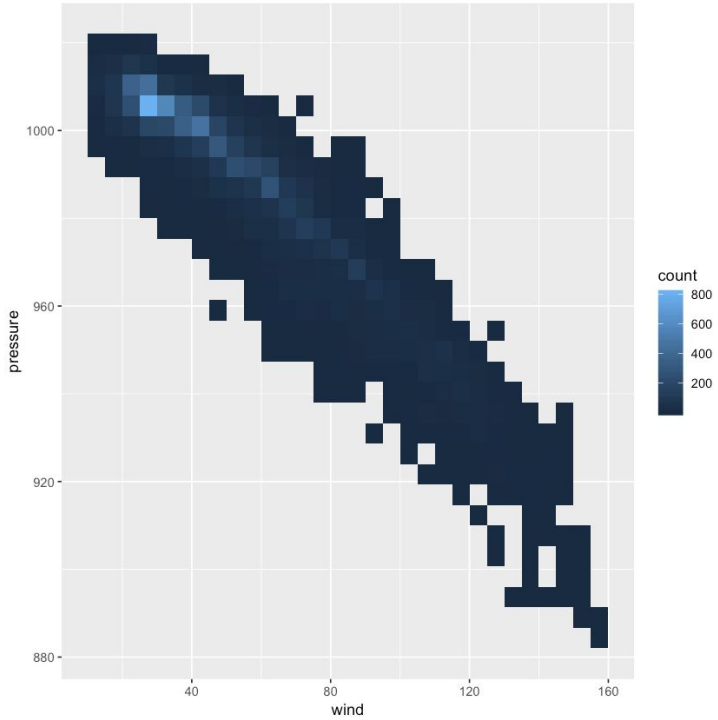
```
In [87]: storms %>%
           ggplot(aes(x = wind, y = pressure)) +
           geom_point() +
           geom_jitter()    # jitter b/c of integer rounding of data
```



Numeric vs. Numeric

2D Histogram
(heatmap)

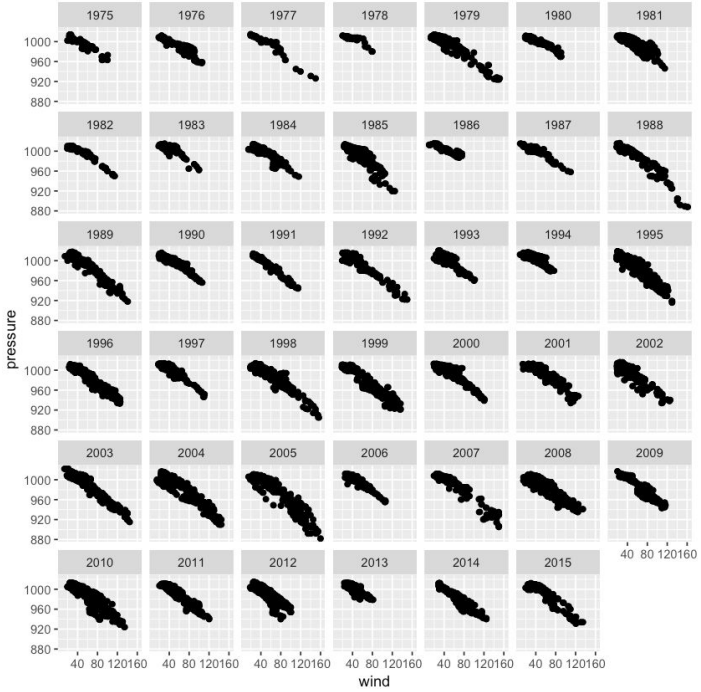
```
In [88]: storms %>%
           ggplot(aes(x = wind, y = pressure)) +
           geom_bin2d()
```



Numeric vs. Numeric vs. Categorical

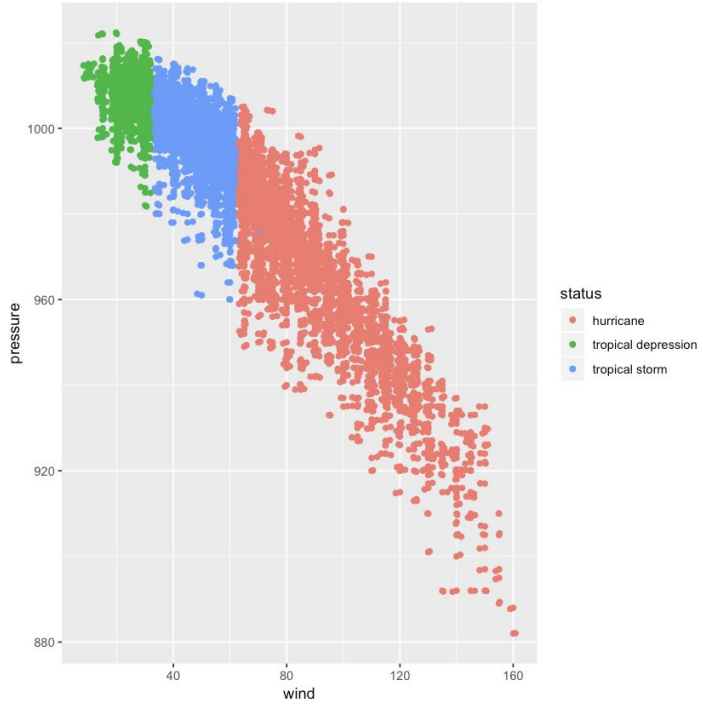
Year (discrete) can be treated as numeric or categorical

```
In [90]: storms %>%
           ggplot(aes(x = wind, y = pressure)) +
           geom_point() +
           geom_jitter() +
           facet_wrap( ~ year)
```



Numeric vs. Numeric vs. Categorical

```
In [92]: storms %>%
           ggplot(aes(x = wind, y = pressure, color = status)) +
           geom_point() +
           geom_jitter()
```



Numeric vs. Categorical

In [94]:
```
diamonds %>%
  group_by(cut) %>%
  summarize(mean = mean(carat), standard_dev = sd(carat))
```

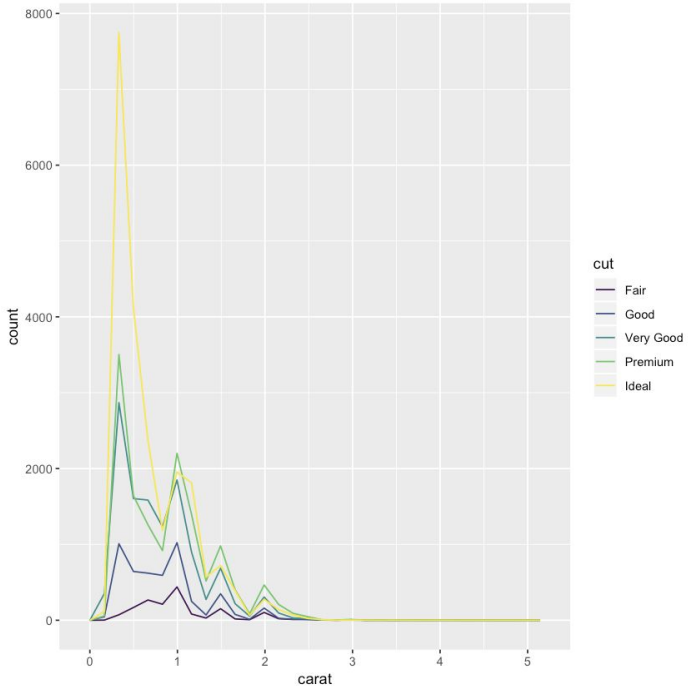
cut	mean	standard_dev
Fair	1.0461366	0.5164043
Good	0.8491847	0.4540544
Very Good	0.8063814	0.4594354
Premium	0.8919549	0.5152616
Ideal	0.7028370	0.4328763

Numeric vs. Categorical

Multiple histograms

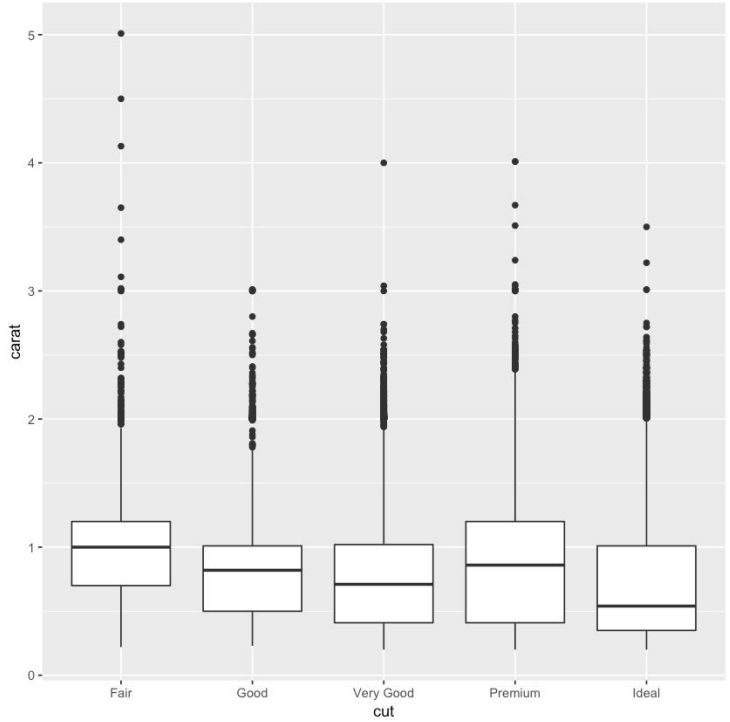
```
In [98]: diamonds %>%
           ggplot(aes(x = carat)) +
           geom_freqpoly(aes(colour = cut))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



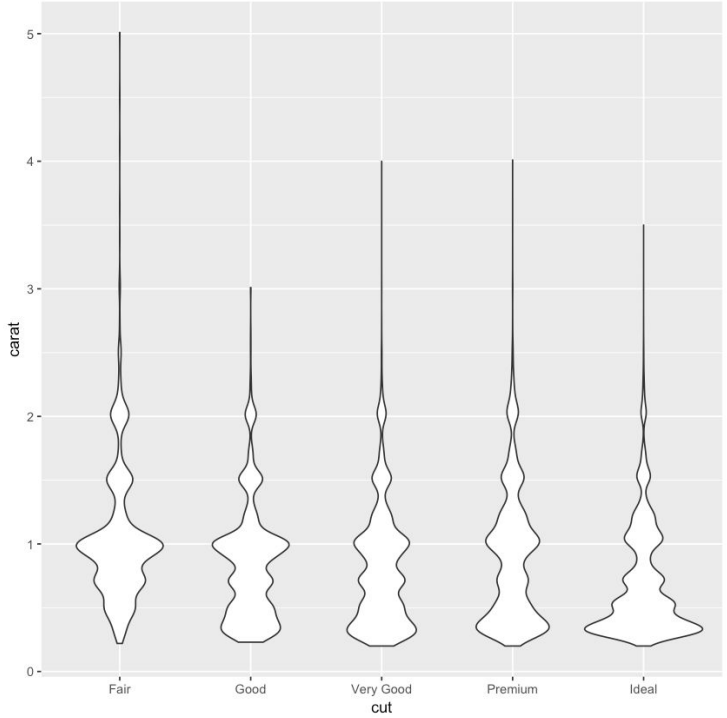
Numeric vs. Categorical

```
In [99]: diamonds %>%
           ggplot(aes(x = cut, y = carat)) +
           geom_boxplot()
```



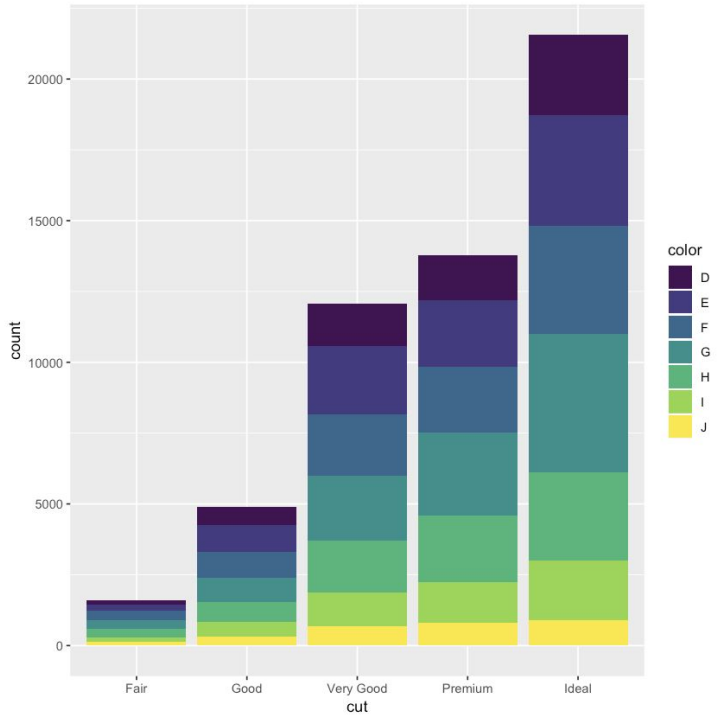
Numeric vs. Categorical

```
In [100]: diamonds %>%
            ggplot(aes(x = cut, y = carat)) +
            geom_violin()
```



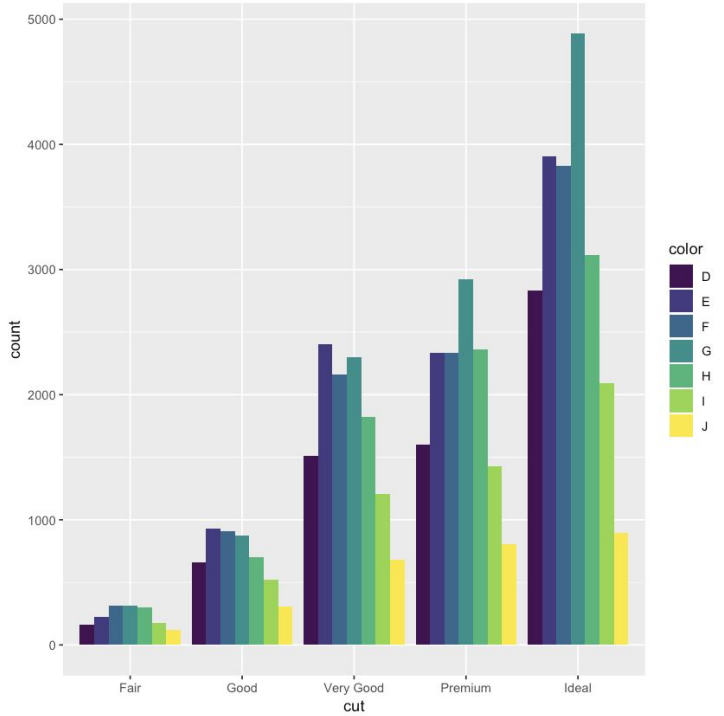
Categorical vs. Categorical

```
In [117]: diamonds %>%
            ggplot(aes(x = cut)) +
            geom_bar(aes(fill = color))
```



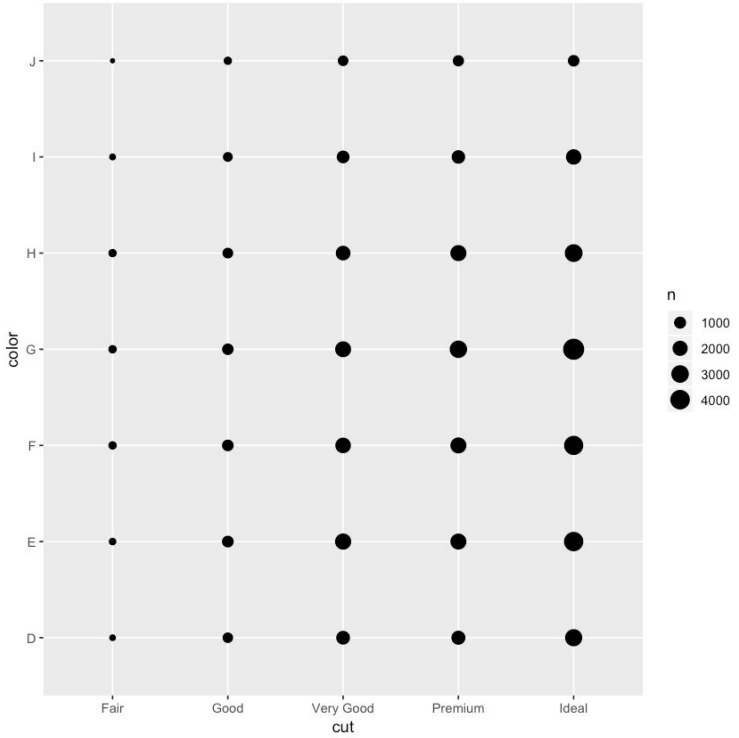
Categorical vs. Categorical

```
In [120]: diamonds %>%
            ggplot(aes(x = cut)) +
            geom_bar(aes(fill = color), position = "dodge")
```



Categorical vs. Categorical

```
In [121]: diamonds %>%
            ggplot() +
            geom_count(mapping = aes(x = cut, y = color))
```



Missing and Uncommon values

- Questions:
  - What? Why?
- Strategies
  - Inspect outliers more closely
  - Remove outliers
  - Turn outliers into Missing Values
  - Do nothing
- Advanced
  - Replace missing values with the mean/mode/median
  - Replace missing values with predictions based on other variables

Summary

- One variable analysis
  - Numeric: Histogram, boxplot, mean, variation
  - Categorical: Counts, bar graph
  - Observe typical vs. uncommon values
- Multivariate analysis
  - Numeric vs. Numeric: scatter plot, heatmap (put into bins and treat as categorical or add aesthetics for other categorical variables)
  - Numeric vs. Categorical: numeric analysis by group
  - Categorical vs. Categorical: cross table, fancy bar graphs