Study Guide Week 1
DATA 110 – Saidi

Intro to DATA 110 – Data Communication and Visualization

"The simple graph has brought more information to the data analyst's mind than any other device." John Tukey

- Intro to Data Vis
- Size of Data
- Telling Stories with Data
- Better Visualizations
- Shape of Data
- Categorical vs Continuous
- Introducing R Studio and R Markdown
- Bar Charts for Categorical Data
- Homework Week 1

Why Data Visualization?

Visualizations tell stories, find hidden patterns in data, and make compelling cases for making decisions. These stories can help you solve real-world problems, like decreasing crime, improving healthcare, and moving traffic on the freeway. They can shed light on shed light on community issues and can help organizations make important decisions.

John Tukey

1. The greatest value of a picture is when it forces us to notice what we never expected to see.
2. The simple graph has brought more information to the data analyst's mind than any other device.
3. Numerical quantities focus on **expected** values, graphical summaries on **unexpected** values.

Data visualization is part art and part science. The challenge is to get the art right without getting the science wrong and vice versa. A data visualization first and foremost has to accurately convey the data. It must not mislead or distort. If one number is twice as large as another, but in the visualization they look to be about the same, then the visualization is wrong. At the same time, a data visualization should be aesthetically pleasing. Good visual presentations tend to enhance the message of the visualization. If a figure contains jarring colors, imbalanced visual elements, or other features that distract, then the viewer will find it harder to inspect the figure and interpret it correctly (Fundamentals of Data Visualization, *Claus O. Wilke)*

A shift in release of government data came in mid-2009 with the launch of Data.gov, a comprehensive catalog of data provided by federal agencies and represents transparency and accountability of groups and officials. Now all this information is in one place and better formatted for analysis and visualization.

A data scientist is someone who **makes value out of data**. They use data to understand and explain the phenomena around them, and help organizations make better decisions.

The Size of Data (in case you didn't know….)

Potential database volumes in bytes for some typical applications (volumes estimated to the nearest order of magnitude). Strictly, bytes are counted in powers of 2 – 1 kilobyte is 1024 bytes, not 1000.		
1 megabyte	1 000 000 (2^{20})	Single data set in a small project database
1 gigabyte	1 000 000 000 (2^{30})	Entire street network of a large city or small country
1 terabyte	1 000 000 000 000 (2^{40})	Elevation of entire Earth surface recorded at 30 m intervals
1 petabyte	1 000 000 000 000 000 (2^{50})	Satellite image of entire Earth surface at 1 m resolution
1 exabyte	1 000 000 000 000 000 000 (2^{60})	A possible 3-D representation of the entire Earth at 10 m resolution
1 zettabyte	1 000 000 000 000 000 000 000 (2^{70})	One-fifth of the capacity (in 2013) of U.S. National Security Agency Utah Data Center

It is not necessarily true that Mark Twain said: "Data is **like garbage. You had better know what you are going to do with it before you collect it**." (but it's a good quote)

Telling Stories with Data

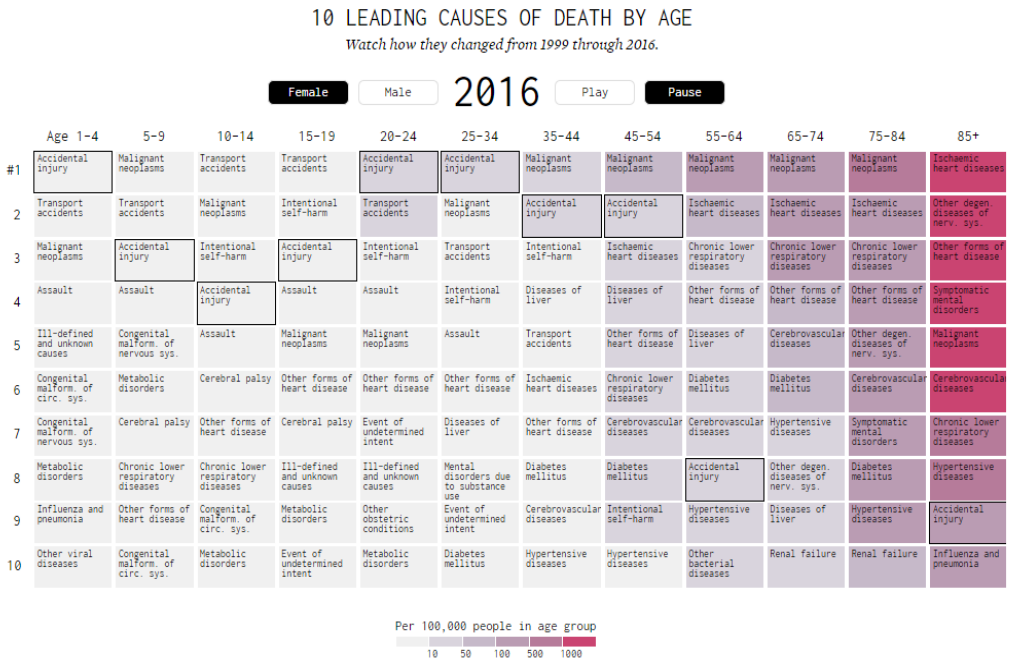
Some Reputable Journalism Sites with Visualizations

- Vox
- New York Times
- Wall Street Journal
- Bloomberg News
- Fivethirtyeight.com
- Washington Post
- Gapminder.org
- The Economist Graphic Detail (https://www.economist.com/graphic-detail/)
- FlowingData (https://flowingdata.com/) by Nathan Yau or go to Data Underload: https://flowingdata.com/category/projects/data-underload/

Intersection of Data, Art and Entertainment

Data can be both informative and artful. Check out Nathan Yau’s many examples on https://flowingdata.com/.

Look at the image of Nathan Yau’s “10 Leading Causes of Death by Age” https://flowingdata.com/2018/10/02/shifting-death/ (see screen capture below).



10 LEADING CAUSES OF DEATH BY AGE

Watch how they changed from 1999 through 2016.

	Age 1-4	5-9	10-14	15-19	20-24	25-34	35-44	45-54	55-64	65-74	75-84	85+
#1	Accidental injury	Malignant neoplasms	Transport accidents	Transport accidents	Accidental injury	Accidental injury	Malignant neoplasms	Malignant neoplasms	Malignant neoplasms	Malignant neoplasms	Malignant neoplasms	Ischaemic heart diseases
2	Transport accidents	Transport accidents	Malignant neoplasms	Intentional self-harm	Transport accidents	Malignant neoplasms	Accidental injury	Accidental injury	Ischaemic heart diseases	Ischaemic heart diseases	Ischaemic heart diseases	Other degen. diseases of nerv. sys.
3	Malignant neoplasms	Accidental injury	Intentional self-harm	Accidental injury	Intentional self-harm	Transport accidents	Intentional self-harm	Ischaemic heart diseases	Chronic lower respiratory diseases	Chronic lower respiratory diseases	Chronic lower respiratory diseases	Other forms of heart disease
4	Assault	Assault	Accidental injury	Assault	Assault	Intentional self-harm	Diseases of liver	Diseases of liver	Other forms of heart disease	Other forms of heart disease	Other forms of heart disease	Symptomatic mental disorders
5	Ill-defined and unknown causes	Congenital malform. of nervous sys.	Assault	Malignant neoplasms	Malignant neoplasms	Assault	Transport accidents	Other forms of heart disease	Diseases of liver	Cerebrovascular diseases	Other degen. diseases of nerv. sys.	Malignant neoplasms
6	Congenital malform. of circ. sys.	Metabolic disorders	Cerebral palsy	Other forms of heart disease	Other forms of heart disease	Other forms of heart disease	Ischaemic heart diseases	Chronic lower respiratory diseases	Diabetes mellitus	Diabetes mellitus	Cerebrovascular diseases	Cerebrovascular diseases
7	Congenital malform. of nervous sys.	Cerebral palsy	Other forms of heart disease	Cerebral palsy	Event of undetermined intent	Diseases of liver	Other forms of heart disease	Cerebrovascular diseases	Cerebrovascular diseases	Hypertensive diseases	Symptomatic mental disorders	Chronic lower respiratory diseases
8	Metabolic disorders	Chronic lower respiratory diseases	Chronic lower respiratory diseases	Ill-defined and unknown causes	Ill-defined and unknown causes	Mental disorders due to substance use	Diabetes mellitus	Diabetes mellitus	Accidental injury	Other degen. diseases of nerv. sys.	Diabetes mellitus	Hypertensive diseases
9	Influenza and pneumonia	Other forms of heart disease	Congenital malform. of circ. sys.	Metabolic disorders	Other obstetric conditions	Event of undetermined intent	Cerebrovascular diseases	Intentional self-harm	Hypertensive diseases	Diseases of liver	Hypertensive diseases	Accidental injury
10	Other viral diseases	Congenital malform. of circ. sys.	Metabolic disorders	Event of undetermined intent	Metabolic disorders	Diabetes mellitus	Hypertensive diseases	Hypertensive diseases	Other bacterial diseases	Renal failure	Renal failure	Influenza and pneumonia

Per 100,000 people in age group

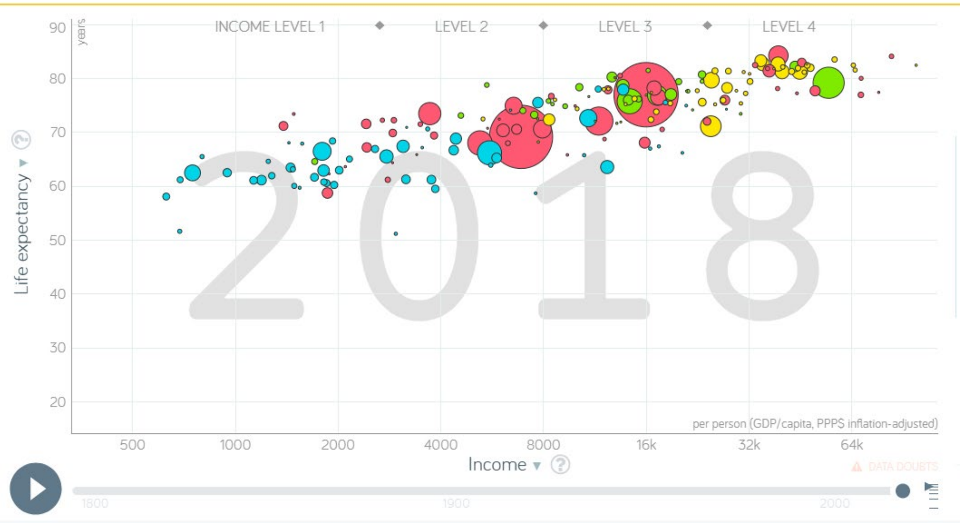
Discuss the following questions:

1. Name all of the variables Yau considers.
2. What do the colors represent?
3. What is Yau’s intended effect in using animation?
4. Do you believe this is an effective data visualization? Why or why not?

Compelling Stories Through Data Exploration

200 Years, 200 Countries, by Hans Rosling

Gapminder.org (You should explore this entire site, but be sure to check out the Tools and Data links)



Better Visualizations

Check out flowingdata.com

When creating visualizations, always consider the following:

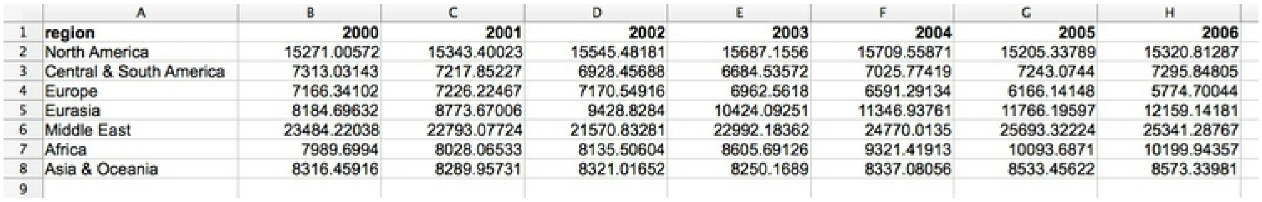
- Patterns
- Relationships
- Design
- Coding (legends and labels)
- Geometrical scaling
- Include source
- Consider your audience

Principles of Analytic Graphics

- Show comparisons
- Show causality, mechanism, explanation
- Show multivariate data
- Integrate multiple modes of evidence
- Describe and document the evidence
- Content is king

What shape is your data?

Particularly when data shows a time series for a single variable, it is often provided like this data on trends in international oil production by region, in "wide" format:

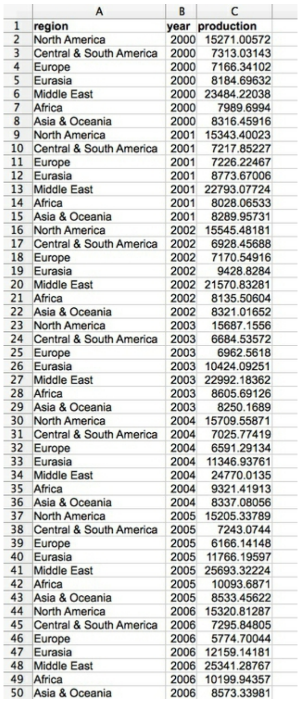


	A	B	C	D	E	F	G	H
1	region	2000	2001	2002	2003	2004	2005	2006
2	North America	15271.00572	15343.40023	15545.48181	15687.1556	15709.55871	15205.33789	15320.81287
3	Central & South America	7313.03143	7217.85227	6928.45688	6684.53572	7025.77419	7243.0744	7295.84805
4	Europe	7166.34102	7226.22467	7170.54916	6962.5618	6591.29134	6166.14148	5774.70044
5	Eurasia	8184.69632	8773.67006	9428.8284	10424.09251	11346.93761	11766.19597	12159.14181
6	Middle East	23484.22038	22793.07724	21570.83281	22992.18362	24770.0135	25693.32224	25341.28767
7	Africa	7989.6994	8028.06533	8135.50604	8605.69126	9321.41913	10093.6871	10199.94357
8	Asia & Oceania	8316.45916	8289.95731	8321.01652	8250.1689	8337.08056	8533.45622	8573.33981
9								

(Source: Peter Aldhous, from U.S. Energy Information Administration data)

Here, all of the numbers represent the same variable, and there is a column for each year. This is good for people to read, but most software for data analysis and visualization does not work well with data in this format.

So if you receive **“wide” data,** you will usually need to convert it to **“long” format (**shown here) using some software, such as R:



	A	B	C
1	region	year	production
2	North America	2000	15271.00572
3	Central & South America	2000	7313.03143
4	Europe	2000	7166.34102
5	Eurasia	2000	8184.69632
6	Middle East	2000	23484.22038
7	Africa	2000	7989.6994
8	Asia & Oceania	2000	8316.45916
9	North America	2001	15343.40023
10	Central & South America	2001	7217.85227
11	Europe	2001	7226.22467
12	Eurasia	2001	8773.67006
13	Middle East	2001	22793.07724
14	Africa	2001	8028.06533
15	Asia & Oceania	2001	8289.95731
16	North America	2002	15545.48181
17	Central & South America	2002	6928.45688
18	Europe	2002	7170.54916
19	Eurasia	2002	9428.8284
20	Middle East	2002	21570.83281
21	Africa	2002	8135.50604
22	Asia & Oceania	2002	8321.01652
23	North America	2003	15687.1556
24	Central & South America	2003	6684.53572
25	Europe	2003	6962.5618
26	Eurasia	2003	10424.09251
27	Middle East	2003	22992.18362
28	Africa	2003	8605.69126
29	Asia & Oceania	2003	8250.1689
30	North America	2004	15709.55871
31	Central & South America	2004	7025.77419
32	Europe	2004	6591.29134
33	Eurasia	2004	11346.93761
34	Middle East	2004	24770.0135
35	Africa	2004	9321.41913
36	Asia & Oceania	2004	8337.08056
37	North America	2005	15205.33789
38	Central & South America	2005	7243.0744
39	Europe	2005	6166.14148
40	Eurasia	2005	11766.19597
41	Middle East	2005	25693.32224
42	Africa	2005	10093.6871
43	Asia & Oceania	2005	8533.45622
44	North America	2006	15320.81287
45	Central & South America	2006	7295.84805
46	Europe	2006	5774.70044
47	Eurasia	2006	12159.14181
48	Middle East	2006	25341.28767
49	Africa	2006	10199.94357
50	Asia & Oceania	2006	8573.33981

(Source: Peter Aldhous, from U.S. Energy Information Administration data)

Notice that now there is one column for each variable, which makes it easier for computers to understand.

Types of data: categorical vs. numerical (from Peter Aldhous)

Before analyzing a dataset, or attempting to draw a graphic, it is important to consider what, exactly, you are working with.

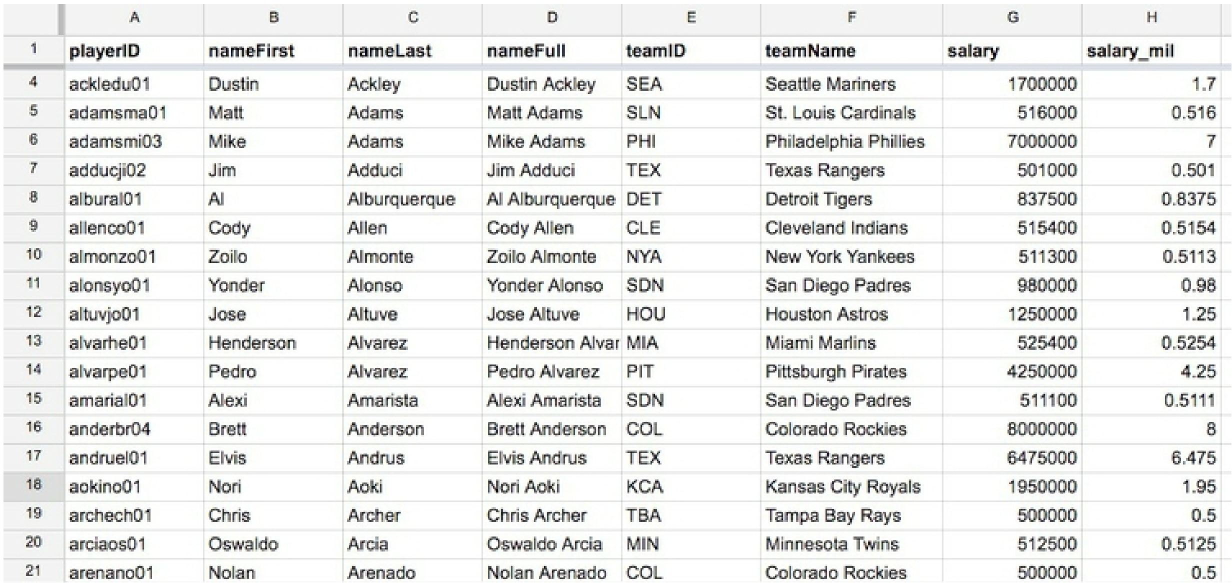
Statisticians often use the term “variable.” This simply means any measure or attribute describing a particular item, or “record,” in a dataset. For example, school students might gather data about themselves for a class project, recording their gender and eye color, and height and weight. There is an important difference between gender and eye color, called “categorical” variables, and height and weight, termed “continuous.”

- **Categorical** variables are descriptive labels given to individual records, assigning them to different groups. The simplest categorical data is dichotomous, meaning that there are just two possible groups — in an election, for instance, people either voted, or they did not. More commonly, there are multiple categories. We analyze categorical data differently than continuous or discrete quantitative data.

- **Numerical** data is richer, consisting of numbers that can have a range of values on a sliding scale. Numerical data may be continuous or discrete (more on this below). Variables might include temperature and amount of rainfall, age, height, weight, blood pressure, etc.

There is a third type of data we often need to consider: **date and time**. One of the most common task in data journalism is to consider how a variable or variables have changed over time.

Datasets will usually contain a mixture of categorical and continuous variables. Here, for example, is a small part of a spreadsheet containing data on salaries for Major League Baseball players at the opening of the 2014 season:



	A	B	C	D	E	F	G	H
1	playerID	nameFirst	nameLast	nameFull	teamID	teamName	salary	salary_mil
4	ackledu01	Dustin	Ackley	Dustin Ackley	SEA	Seattle Mariners	1700000	1.7
5	adamsma01	Matt	Adams	Matt Adams	SLN	St. Louis Cardinals	516000	0.516
6	adamsmi03	Mike	Adams	Mike Adams	PHI	Philadelphia Phillies	7000000	7
7	adducji02	Jim	Adduci	Jim Adduci	TEX	Texas Rangers	501000	0.501
8	albural01	Al	Alburquerque	Al Alburquerque	DET	Detroit Tigers	837500	0.8375
9	allenco01	Cody	Allen	Cody Allen	CLE	Cleveland Indians	515400	0.5154
10	almonzo01	Zoilo	Almonte	Zoilo Almonte	NYA	New York Yankees	511300	0.5113
11	alonsyo01	Yonder	Alonso	Yonder Alonso	SDN	San Diego Padres	980000	0.98
12	altuvjo01	Jose	Altuve	Jose Altuve	HOU	Houston Astros	1250000	1.25
13	alvarhe01	Henderson	Alvarez	Henderson Alvar	MIA	Miami Marlins	525400	0.5254
14	alvarpe01	Pedro	Alvarez	Pedro Alvarez	PIT	Pittsburgh Pirates	4250000	4.25
15	amarial01	Alexi	Amarista	Alexi Amarista	SDN	San Diego Padres	511100	0.5111
16	anderbr04	Brett	Anderson	Brett Anderson	COL	Colorado Rockies	8000000	8
17	andruel01	Elvis	Andrus	Elvis Andrus	TEX	Texas Rangers	6475000	6.475
18	aokino01	Nori	Aoki	Nori Aoki	KCA	Kansas City Royals	1950000	1.95
19	archech01	Chris	Archer	Chris Archer	TBA	Tampa Bay Rays	500000	0.5
20	arciaos01	Oswaldo	Arcia	Oswaldo Arcia	MIN	Minnesota Twins	512500	0.5125
21	arenano01	Nolan	Arenado	Nolan Arenado	COL	Colorado Rockies	500000	0.5

(Source: Peter Aldhous, data from Lahman Baseball Database data)

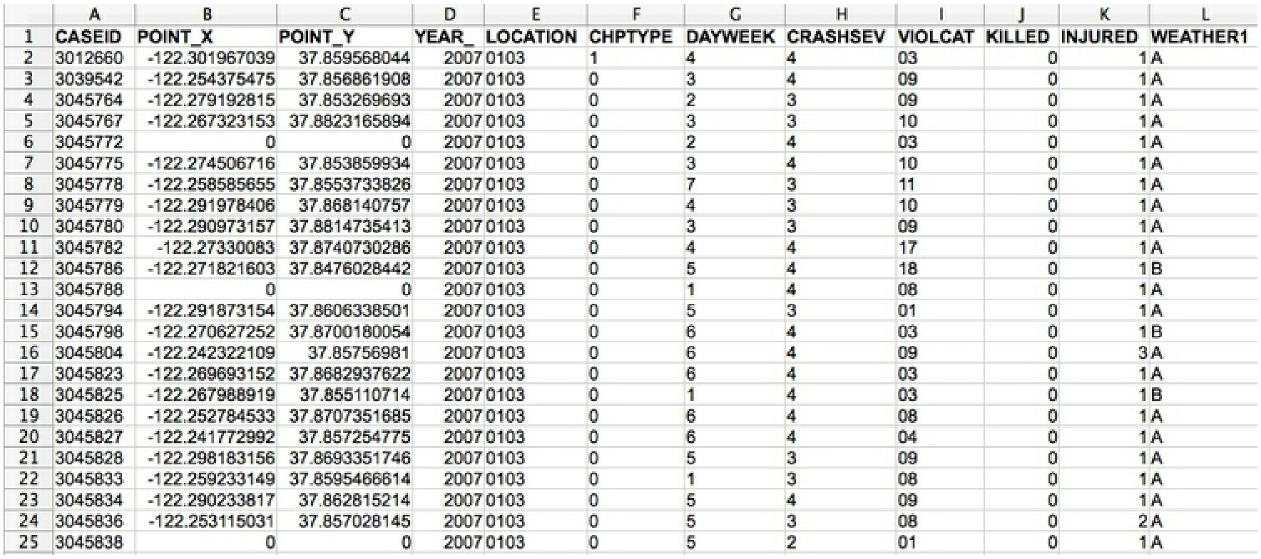
This is a typical data table layout

- the players --- form the rows
- variables --- arranged in columns

It is easy to recognize the categorical variables of teamID and teamName because they are each entered as text. The numbers for salary, expressed in full or in millions of dollars (salary_mil), are continuous variables.

Never assume that every **number** in a dataset represents a **continuous variable.** Text descriptions can make datasets unwieldy, so database managers often adopt simpler codes such as letting numbers store categorical data.

You can see this in the following example, showing data on traffic accidents resulting in injury or death in Berkeley, downloaded from a database maintained by researchers on the Berkeley campus.



	A	B	C	D	E	F	G	H	I	J	K	L
1	CASEID	POINT_X	POINT_Y	YEAR_	LOCATION	CHPTYPE	DAYWEEK	CRASHSEV	VIOLCAT	KILLED	INJURED	WEATHER1
2	3012660	-122.301967039	37.859568044	2007	0103	1	4	4	03	0	1	A
3	3039542	-122.254375475	37.856861908	2007	0103	0	3	4	09	0	1	A
4	3045764	-122.279192815	37.853269693	2007	0103	0	2	3	09	0	1	A
5	3045767	-122.267323153	37.8823165894	2007	0103	0	3	3	10	0	1	A
6	3045772	0	0	2007	0103	0	2	4	03	0	1	A
7	3045775	-122.274506716	37.853859934	2007	0103	0	3	4	10	0	1	A
8	3045778	-122.258585655	37.8553733826	2007	0103	0	7	3	11	0	1	A
9	3045779	-122.291978406	37.868140757	2007	0103	0	4	3	10	0	1	A
10	3045780	-122.290973157	37.8814735413	2007	0103	0	3	3	09	0	1	A
11	3045782	-122.27330083	37.8740730286	2007	0103	0	4	4	17	0	1	A
12	3045786	-122.271821603	37.8476028442	2007	0103	0	5	4	18	0	1	B
13	3045788	0	0	2007	0103	0	1	4	08	0	1	A
14	3045794	-122.291873154	37.8606338501	2007	0103	0	5	3	01	0	1	A
15	3045798	-122.270627252	37.8700180054	2007	0103	0	6	4	03	0	1	B
16	3045804	-122.242322109	37.85756981	2007	0103	0	6	4	09	0	3	A
17	3045823	-122.269693152	37.8682937622	2007	0103	0	6	4	03	0	1	A
18	3045825	-122.267988919	37.855110714	2007	0103	0	1	4	03	0	1	B
19	3045826	-122.252784533	37.8707351685	2007	0103	0	6	4	08	0	1	A
20	3045827	-122.241772992	37.857254775	2007	0103	0	6	4	04	0	1	A
21	3045828	-122.298183156	37.8693351746	2007	0103	0	5	3	09	0	1	A
22	3045833	-122.259233149	37.8595466614	2007	0103	0	1	3	08	0	1	A
23	3045834	-122.290233817	37.862815214	2007	0103	0	5	4	09	0	1	A
24	3045836	-122.253115031	37.857028145	2007	0103	0	5	3	08	0	2	A
25	3045838	0	0	2007	0103	0	5	2	01	0	1	A

(Source: Peter Aldhous, from Transportation Injury Mapping System data)

Of the numbers seen here, only the YEAR, latitudes and longitudes (POINT_Y and POINT_X) and numbers of people KILLED or INJURED actually represent continuous variables. (Look carefully, and you will see that these **numbers are justified right within each cell**. **The other numbers are justified left**, like the text entries, because they were imported into the spreadsheet as text values.)

Like this example, many datasets are difficult to interpret without their supporting documentation. So each time you acquire a dataset, if necessary make sure you also obtain the "codebook" describing all of the variables/fields, and how they are coded. Here is the codebook for the traffic accident data.

Working with categorical data

You might imagine that there is little that you can do with categorical data alone, but it can be powerful, and can also be used to create new continuous variables.

The most basic operation with categorical data is to aggregate it by counting the number of records that fall into each category. This gives a table of **frequencies.** These are often divided by the total number of records, then multiplied by 100 to show them as percentages of the total.

Creating frequency counts from categorical data creates a new continuous variable — what has changed is the level of analysis. In this example, the original data would consist of a huge table with a record for each person, noting their racial/ethnic identity as categorical variables; in creating the frequency table shown here, the level of analysis has shifted from the individual to the racial/ethnic group.

We can ask more interesting questions by considering two categorical variables together.

NY Stop and Frisk Program Data

Below is 2011 random sample data of police stops from New York's "Stop and Frisk" program. The program allowed police officers to stop people on the street and search them for weapons or contraband. The program was controversial. Critics alleged that it led to heightened police discrimination of minorities. The results are combined into a "contingency table" or "cross-tab":

	Race of Suspect			
Level of Force	***Black***	***Hispanic***	***White***	***Marginal Totals***
No force used	*971*	*853*	*260*	*2084*
Hands used	*263*	*188*	*30*	*481*
Higher force level *	*109*	*72*	*18*	*199*
Marginal Totals	*1343*	*1113*	*308*	*2764*

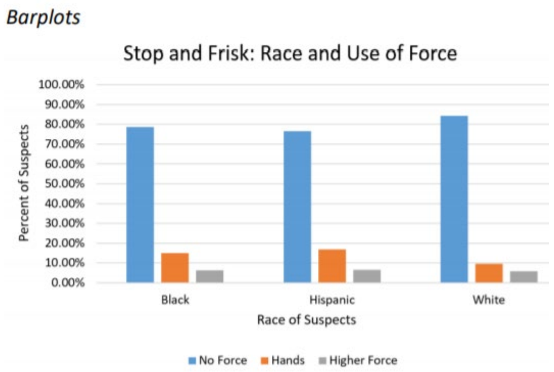
* Includes push to wall/ground, handcuffs, draw/point weapon, pepper spray, baton
Data source: NYC.gov https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

We can see that there are 2 categorical variables: Level of Force and Race of Suspect

If we just compare counts, we could say that a much higher number of Black suspects than White suspects had interactions with police with no use of force.

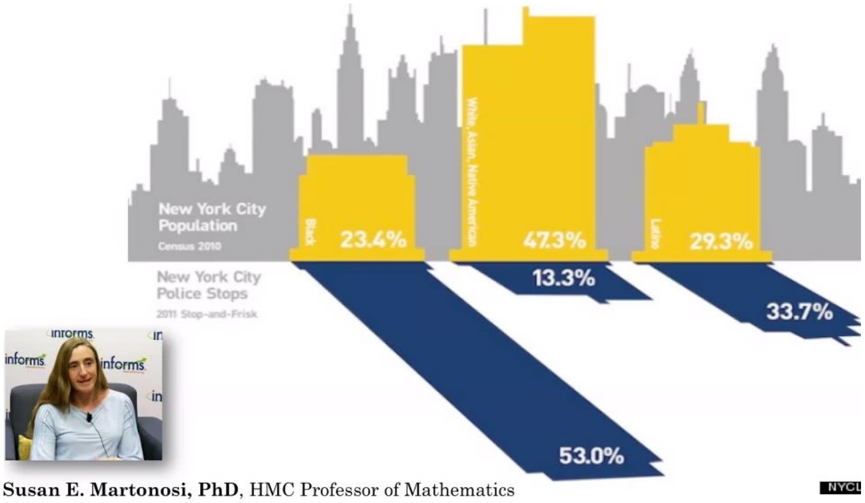
Is that a good analysis?

What about the barplot below?



While the numbers are technically correct, this is not a good analysis of this dataset. The marginal distributions show more black people were stopped than white people. So, if force rates were equal, we'd expect a higher number of no-force stops with black suspects than white suspects. In addition, the proportion of white people who received no force was actually higher than it was among black people. In other words, white people had no-force interactions at higher rates than black people. It's just that fewer white people were stopped and searched in the first place.

Generally, focusing on raw numbers rather than rates/proportions in a two-way table can lead to misleading conclusions.



Susan E. Martonosi, PhD, HMC Professor of Mathematics

Instead, look at the RATES for different races as compared to the overall population proportions.

(The rates in this graph are from more recent data)

Here is more information on An Empirical Analysis of Racial Differences in Police Use of Force (Freyer, 2017) https://scholar.harvard.edu/files/fryer/files/empirical_analysis_tables_figures.pdf

Now on to Numerical data

Numerical variables can be further categorized as **continuous or discrete**.

- **Continuous numerical** variables are usually measured, such as height. These variables can take on an infinite number of values within a given range.

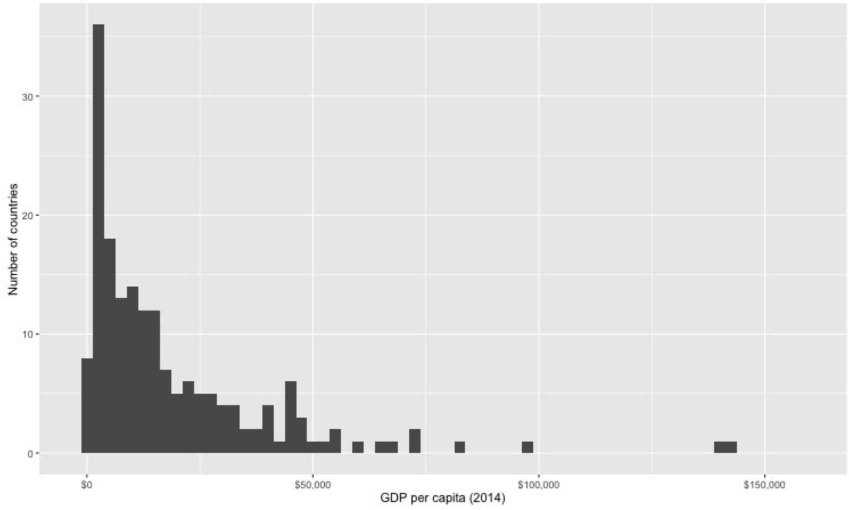
- **Discrete numerical** variables are those that take on one of a specific set of numeric values where we are able to count or enumerate all of the possibilities. One example of a discrete variable is number of pets in a household. In general, count data are an example of discrete variables.

When determining whether a numerical variable is continuous or discrete, it is important to think about the nature of the variable and not just the observed value, as rounding of continuous variables can make them appear to be discrete. For example, height is a continuous variable, however we tend to report our height rounded to the nearest unit of measure, like inches or centimeters.

To produce informative graphics that tell a clear story, data journalists often need to turn a continuous variable into a categorical variable by dividing it into bins.

Selecting the range of the bins depends on the story you are telling. In the jobless rate example, the bins divided the population into groups of young, mid-career and older workers, revealing how young workers in particular were bearing the brunt of the Great Recession. When binning data, it is again a good idea to look at the distribution, and experiment with different possibilities.

For example, the wealth of nations, measured in terms of gross domestic product (GDP) per capita in 2014, has a skewed distribution. If we look at the distribution, drawn in increments of $2,500, we will see that it is highly skewed.



(Source: Peter Aldhous, from World Bank data)

Just a few countries had a GDP per capita of more than 50,000.

You may also want to examine histograms for obvious “valleys” in the data, which may be good places for the breaks between bins.

Explore the **High School and Beyond Survey Data** from the National Center of Education Statistics.

Below is a preview of the dataset. View the variables and the row values in the dataset. You can access the **hsb2** dataset here.

id	gender	race	ses	schtyp	prog	read	write	math	science
<int>	<chr>	<chr>	<fct>	<fct>	<fct>	<int>	<int>	<int>	<int>
70	male	white	low	public	general	57	52	41	47
121	female	white	middle	public	vocational	68	59	53	63
86	male	white	high	public	general	44	33	54	58
141	male	white	high	public	vocational	63	44	47	53
172	male	white	middle	public	academic	47	52	57	53
113	male	white	middle	public	academic	44	52	51	63
50	male	african american	middle	public	general	50	59	42	53
11	male	hispanic	middle	public	academic	34	46	45	39
84	male	white	middle	public	general	63	57	54	58
48	male	african american	middle	public	academic	57	55	52	50

Later, you will learn to load a CSV file that has been saved on your computer into R Studio. For now, the dataset is contained in the **openintro** library, so you will not need to load the dataset at this time.

Take a peek

When you want to work with data in R, a good first step is to take a peek at what the data look like. The head() function is one good way of doing this.

The first variable is id, which is an identifier variable for the student.

```
## Rows: 200
## Columns: 1
## $ id <int> 70, 121, 86, 141, 172, 113, 50, 11, 84, 48, 75, 60, 95, 104, 38, 11…
```

Strictly speaking, this is a categorical variable, though the labeling of this variable is likely not that useful since we would not use this variable in an analysis of relationships between the variables in the dataset. You can think of this variable as being an anonymized version to having the names of the students in the dataset.

The next variable is gender, a categorical variable, with levels `"male"` and `"female"`. It should be noted that the language of government surveys, such as High School and Beyond, is slow to change. So with these types of data, you will continue to see variables mislabeled as “gender” when they in fact measure the biological sex (male, female) of the participant.

There is no inherent ordering to the levels of this variable, no matter what anyone tells you! So, this is just a categorical variable. The same is true for the race variable, which has levels of `"white"`, `"african american"`, `"hispanic"`, and `"asian"`.

```
## Rows: 200
## Columns: 2
## $ race   <chr> "white", "white", "white", "white", "white", "white", "african …
## $ gender <chr> "male", "female", "male", "male", "male", "male", "male", "male…
```

Socio-economic status, on the other hand, has three levels `"low"`, `"middle"`, and `"high"` that have an inherent ordering, hence this variable is an *ordinal* categorical variable.

```
## Rows: 200
## Columns: 1
## $ ses <fct> low, middle, high, high, middle, middle, middle, middle, middle, m…
```

School type and program are also both categorical variables, with no inherent ordering to their levels.

```
## Rows: 200
## Columns: 2
## $ schtyp <fct> public, public, public, public, public, public, public, public,…
## $ prog   <fct> general, vocational, general, vocational, academic, academic, g…
```

The remaining variables are scores that these students received in reading, writing, math, science, and social studies tests. Since these scores are all whole numbers, and assuming that it is not possible to obtain a non-whole number score on these tests, these variables are discrete numerical.

Introducing R and R Studio

For homework, you will install R and R Studio using provided documents.

First, install R from CRAN R Project (https://cran.r-project.org/). Select the version you will need based on your computer type.

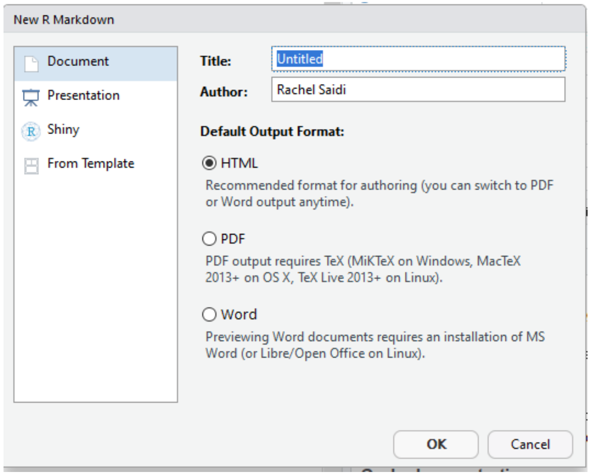
Then install R Studio (https://rstudio.com/products/rstudio/download/). Be sure to select the FREE version.

This is a really useful document with R Markdown Reference Material

R Markdown to Explore Relationships

We will learn to code within the context of using R Markdown. Click here for more information on R Markdown and Quarto

Start by opening a new R Quarto file. Select output format to be HTML or Word or PDF.



Title: Untitled

Author: Rachel Saidi

Default Output Format:

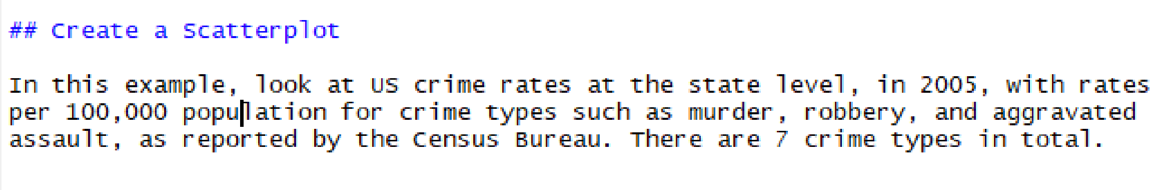
● HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

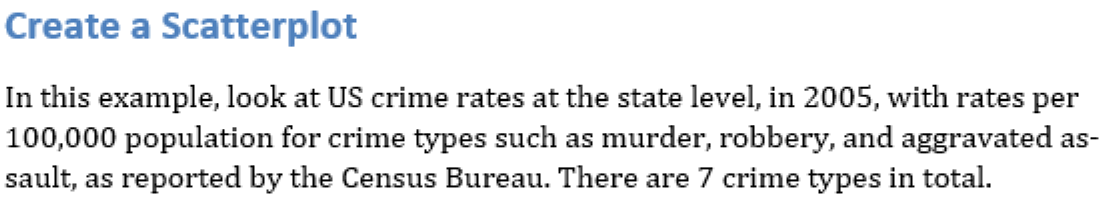
In the white space on the Quarto page, ## will give you subtitles – you know you have done it correctly because they are in blue

When you do not include hashtags, this gives text (in black). You can use #, ##, ###, etc. for different levels of subtitles, or no hashtags at all for plain text.



```
## Create a Scatterplot

In this example, look at US crime rates at the state level, in 2005, with rates
per 100,000 population for crime types such as murder, robbery, and aggravated
assault, as reported by the Census Bureau. There are 7 crime types in total.
```

Here is how it looks when it is knitted in Word:



Create a Scatterplot

In this example, look at US crime rates at the state level, in 2005, with rates per 100,000 population for crime types such as murder, robbery, and aggravated assault, as reported by the Census Bureau. There are 7 crime types in total.

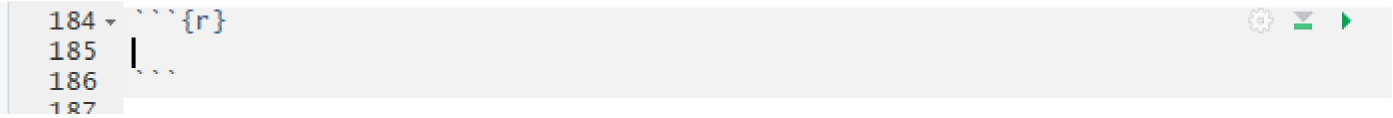
Here is how it looks when knitted to HTML

Create a Scatterplot

In this example, look at US crime rates at the state level, in 2005, with rates per 100,000 population for crime types such as murder, robbery, and aggravated assault, as reported by the Census Bureau. There are 7 crime types in total. The dataset is clean to begin with.

Create Chunks

You can create chunks of code with the shortcut control+ alt+ I. Here is a chunk:



~~~
```{r}

```
~~~

Once you have code typed in, to run just that chunk, click on the right green arrow.

Be sure to leave a line or two of space after a chunk before you add your next subtitle

## Subtitle

If you don't leave a space, it may not knit properly.

Once you have written all your subtitles, notes, and chunks, run the code to be sure it works.

Packages

Packages are the fundamental units of reproducible R code. They include reusable functions, the documentation that describes how to use them, and sample data. In this lesson we will make use of two packages:

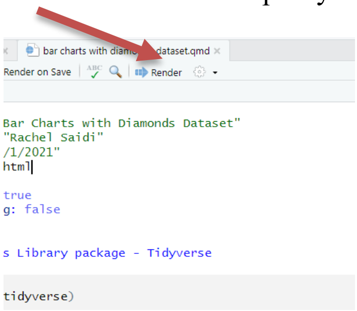
- **tidyverse**: Tidyverse is a collection of R packages for data science that adhere to a common philosophy of data and R programming syntax, and are designed to work together naturally. You can learn more about tidyverse here. But no need to go digging through the package documentation, we will walk you through what you need to know about these packages as they become relevant.

Once we have installed the packages, we use the `library()` function to load packages into R.Let's load these this packages to be used in the remainder of this lesson.

```
library(tidyverse)
```

Rendering

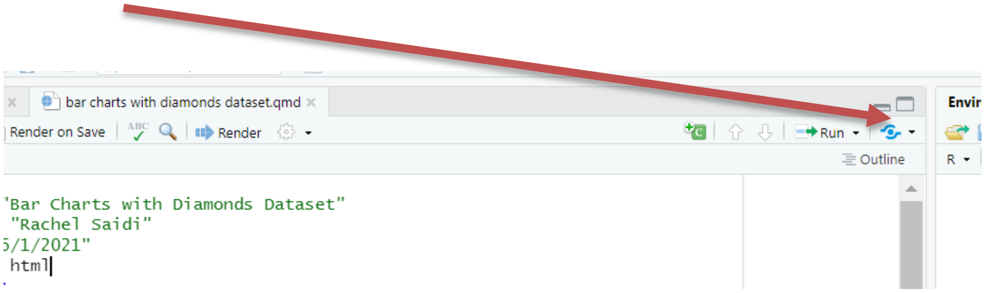
You can **render** at any time to test how it looks. You can switch from knitting to Word, HTML, pdf at any time using the render button at the top of your document code.



```
Bar Charts with Diamonds Dataset"
"Rachel Saidi"
/1/2021"
html

true
g: false


s Library package - Tidyverse

tidyverse)
```

Finally, Publish to Rpubs

When you are done with your Markdown code, or when you just want to see what it looks like knitted to HTML, click the Rpubs “Publish” symbol:



```
'Bar Charts with Diamonds Dataset"
 "Rachel Saidi"
5/1/2021"
 html
```

You will need to set up a free Rpubs account, but once you do so, you will be able to publish your Markdown as an HTML page and share the link. Alternatively, you can publish your work to Github. We will learn about that later in the class.

Bar Charts for Categorical Data (from Nicole Radziwill)

To create a bar chart/bar plot in R with the barplot function using categorical data, which is a collection of numbers that represent frequencies (or counts) of events or outcomes that fall into different groups or categories. [Note: If you are trying to display distributions of quantitative data, choose a histogram instead. Bar charts are for categorical data only. BAR CHARTS ARE NOT THE SAME AS HISTOGRAMS!]

The lengths of the bars are proportional to the values they represent, and the bars can be oriented vertically or horizontally.

- Good bar charts are labeled nicely, with a clear description of the categories that are being counted on the horizontal (x) axis, and a label on the vertical (y) axis that indicates whether frequencies or counts are displayed.

- Membership into each category should be mutually exclusive. That is, you don’t want an observation to appear in multiple bars.

- Pie charts should be avoided. A bar chart is a better way to display your data. If you are trying to illustrate a collection of items that naturally add up to 100%, a pie chart may be appropriate. However, if there are multiple categories where it may be difficult to distinguish which slice is bigger (such as one observation of 28% and another observation of 29%) a bar chart may be more appropriate.

- If you want to display your data in terms of TWO categorical variables, choose a segmented bar chart (described in a separate chapter). Even More Caution: There is a BIG DIFFERENCE between a bar chart and a histogram! Even though a bar chart looks really similar to a histogram at first glance, take a close look at what kind of data is on the horizontal (x) axis. In a bar chart, the horizontal axis lists categories. In a histogram, the horizontal axis will contain ranges of numbers that represent a continuum (e.g. 0-10, 10-20, 20-30 and so forth). Also, in a bar chart, there will be some space between the bars indicating that the categories are separate from one another - whereas with a histogram, there will be no space between the bars! The bars will be very cozy in a histogram, mashed up against one another like they're at a crowded party, whereas the bars in a bar chart need a little more breathing room, and thus are distanced from one another.

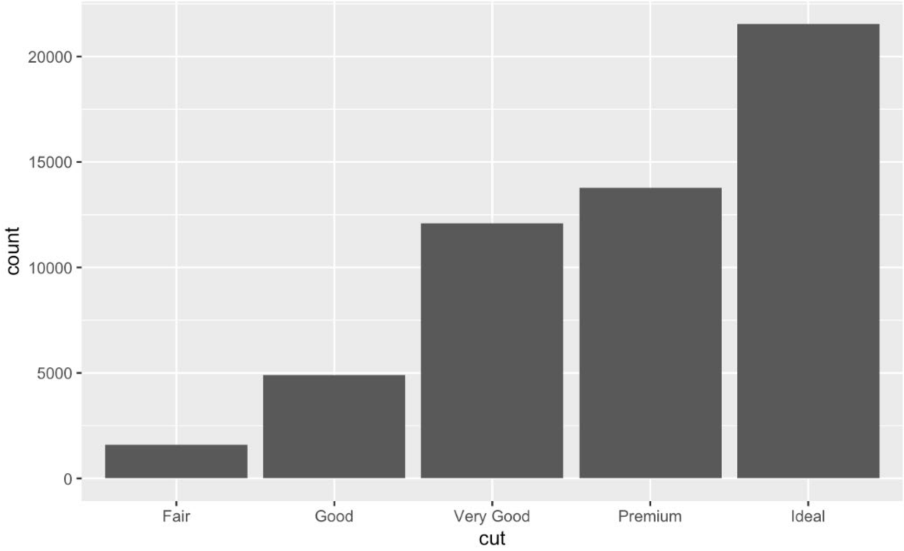
Basic Barplots

https://r4ds.had.co.nz/data-visualisation.html

3.7 Statistical transformations

Next, let's take a look at a bar chart. Bar charts seem simple, but they are interesting because they reveal something subtle about plots. Consider a basic bar chart, as drawn with `geom_bar()`. The following chart displays the total number of diamonds in the `diamonds` dataset, grouped by `cut`. The `diamonds` dataset comes in ggplot2 and contains information about ~54,000 diamonds, including the `price`, `carat`, `color`, `clarity`, and `cut` of each diamond. The chart shows that more diamonds are available with high quality cuts than with low quality cuts.

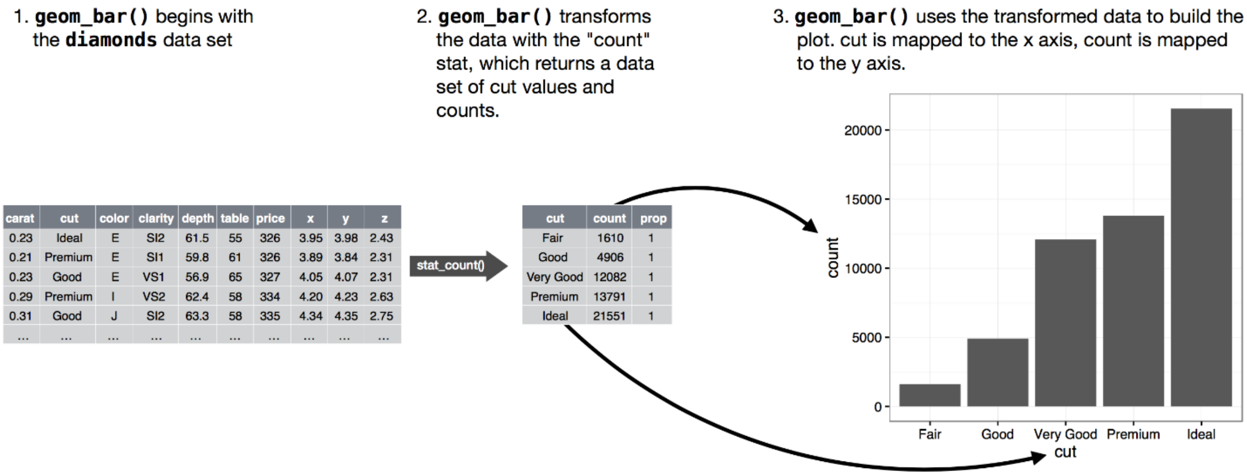
```
ggplot(data = diamonds) + 
  geom_bar(mapping = aes(x = cut))
```



On the x-axis, the chart displays `cut`, a variable from `diamonds`. On the y-axis, it displays count, but count is not a variable in `diamonds`! Where does count come from? Many graphs, like scatterplots, plot the raw values of your dataset. Other graphs, like bar charts, calculate new values to plot:

- bar charts, histograms, and frequency polygons bin your data and then plot bin counts, the number of points that fall in each bin.
- smoothers fit a model to your data and then plot predictions from the model.
- boxplots compute a robust summary of the distribution and then display a specially formatted box.

The algorithm used to calculate new values for a graph is called a **stat**, short for statistical transformation. The figure below describes how this process works with `geom_bar()`.



1. **geom_bar()** begins with the **diamonds** data set

2. **geom_bar()** transforms the data with the "count" stat, which returns a data set of cut values and counts.

3. **geom_bar()** uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

Bar Charts with Diamonds Dataset

Rachel Saidi

2021-06-01

Access Library package - Tidyverse

```
library(tidyverse)
```

Load the pre-built dataset, Diamonds, and view it in the global environment

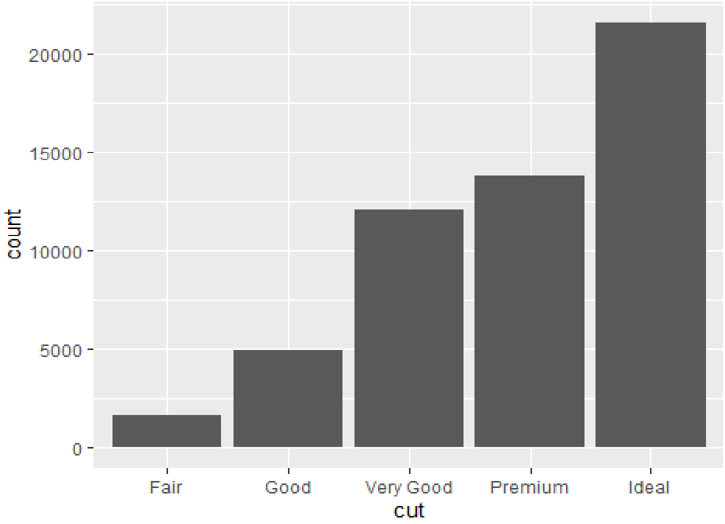
```
head(diamonds)  # shows the first few lines of the dataset

# A tibble: 6 × 10
  carat cut       color clarity depth table price     x     y     z
  <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48

data(diamonds)   # places the dataset in the global environment
```

Statistical transformations (from R for Data Science)

Bar charts seem simple, but they are interesting because they reveal something subtle about plots. Consider a basic bar chart, as drawn with **geom_bar()**. The following chart displays the total number of diamonds in the diamonds dataset, **grouped by cut.** The diamonds dataset comes in ggplot2 and contains information about ~54,000 diamonds, including the **price, carat, color, clarity, and cut** of each diamond. The bar graph shows that more diamonds are available with high quality cuts than with low quality cuts.

First Bar Plot

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



How do bar charts work with 2 variables?

Bar graphs are EASY when you have a single categorical variable that defines several levels for each observation. Ex: “cut” has levels: fair, good, very good, premium, and ideal. Each observation is categorized this way. But what if you have a table of aggregated data: x = cut vs y = frequency?

Here is a tibble to show this table and how you can create a bar graph from this data

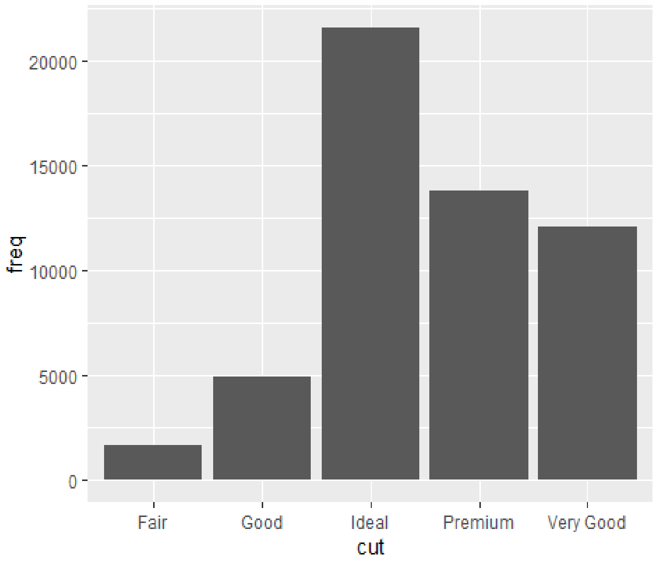
A Tibble (think of this like a dataframe)

We will create a frequency table of the types of cuts that mimick the calculations done to create geom_bar

```
demo <- tribble(
      ~cut,         ~freq,
      "Fair",       1610,
      "Good",       4906,
      "Very Good",  12082,
      "Premium",    13791,
      "Ideal",      21551
    )
```

Demo Tibble Bar Plot Looks just like our other bar graphs

```
ggplot(data = demo) +
  geom_bar(mapping = aes(x = cut, y = freq), stat = "identity")
```



(Don't worry that you haven't seen <- or tribble() before. You might be able to guess at their meaning from the context, and you'll learn exactly what they do soon!)

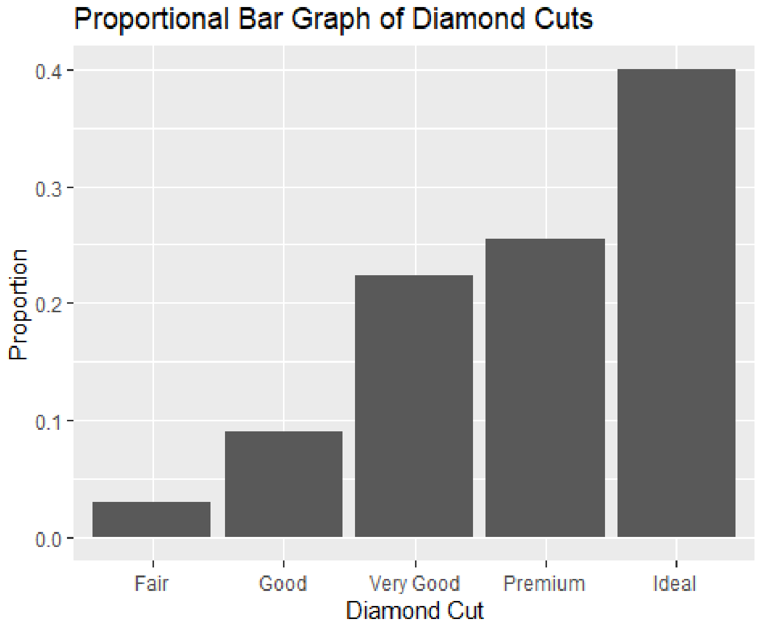
Creating Proportional Bars

You might want to override the default mapping from transformed variables to aesthetics. For example, you might want to display a bar chart of proportion, rather than count:

Proportional Bar Graph (Relative Frequencies)

You need "group=1" when plotting proportions (try to omit it and see)

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, y = stat(prop), group = 1)) +
  labs(x = "Diamond Cut", y = "Proportion",
       title = "Proportional Bar Graph of Diamond Cuts")
```



To find the variables computed by the stat, look for the help section titled “computed variables”.

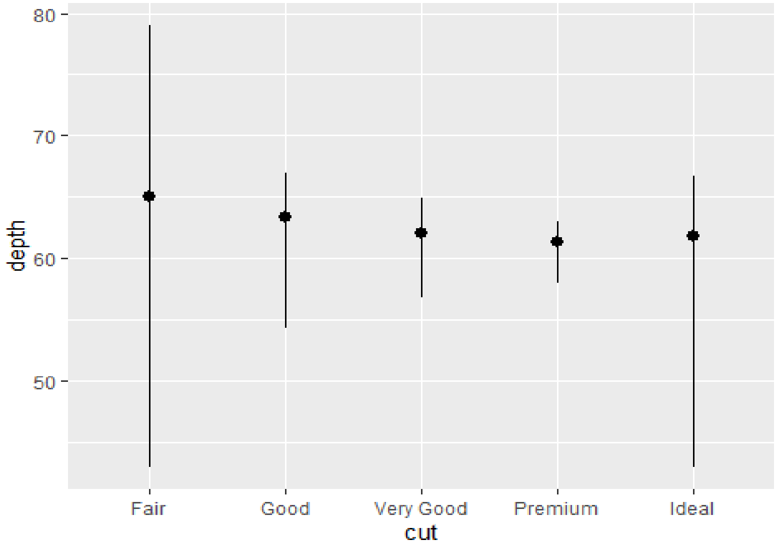
This is a different type of plot that shows a line with min, max, and median values

You might want to draw greater attention to the statistical transformation in your code. For example, you might use stat_summary(), which summarises the y values for each unique x value, to draw attention to the summary that you’re computing:

Line Plot

This is a different way of visualizing center and spread of cuts and depth

```r
ggplot(data = diamonds) +
     stat_summary(
        mapping = aes(x = cut, y = depth),
        fun.min = min,
        fun.max = max,
        fun = median
  )
```



Fill vs Color

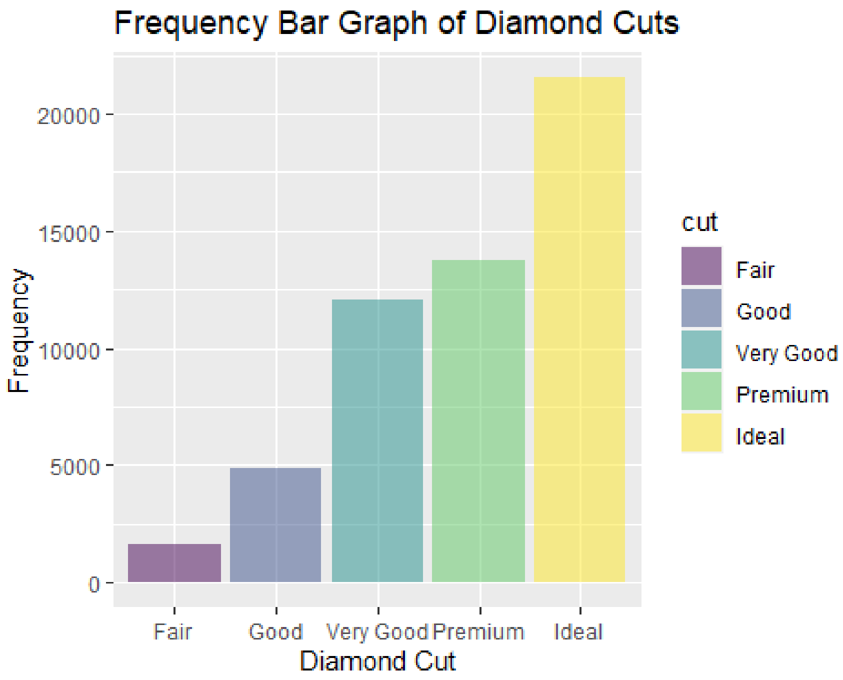
Position adjustments

There’s one more piece of magic associated with bar charts. You can color a bar chart using either the color aesthetic, or, more usefully, fill:

Notice that “fill=” fills the inside of the bar, whereas “color=” draws a color outline of the bar. Alpha gives a level of transparency, with alpha = 0 is invisible and alpha = 1 is fully saturated

Bar Plot with Alpha Transparency

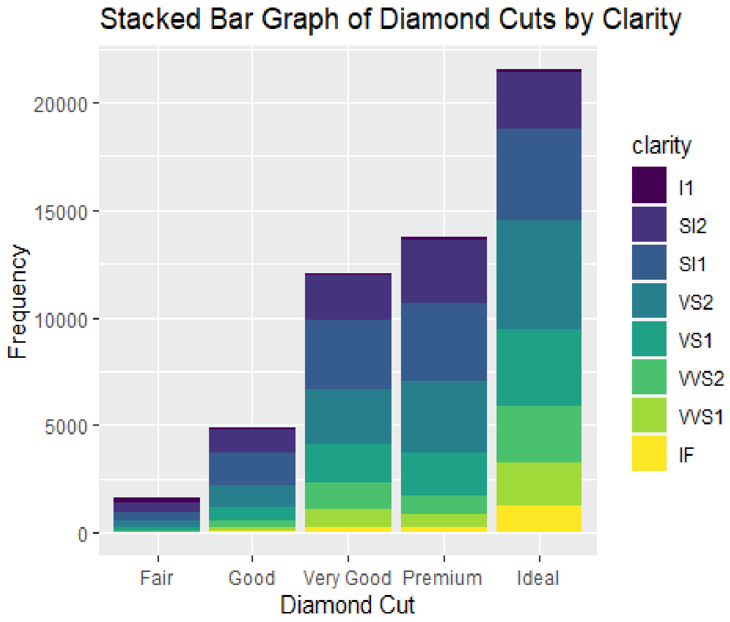
```r
ggplot(data = diamonds, aes(x=cut, fill = cut)) +
  geom_bar(alpha = 0.5)+  # try replacing alpha = 0.5 with 0.8 to see how it changes
  labs(x = "Diamond Cut", y = "Frequency",
       title = "Frequency Bar Graph of Diamond Cuts")
```



Try stacking bar graphs with position = "stack"

Note what happens if you map the fill aesthetic to another variable, like clarity: the bars are automatically stacked. Each colored rectangle represents a combination of cut and clarity.

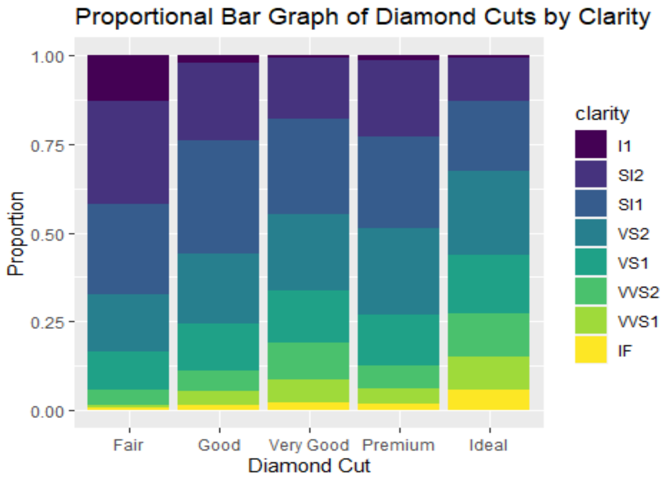
```
ggplot(data = diamonds) +
  geom_bar(aes(x = cut, fill = clarity), position = "stack") +
    labs(x = "Diamond Cut",
         y = "Frequency",
         title = "Stacked Bar Graph of Diamond Cuts by Clarity")
```



The identity position adjustment is more useful for 2d geoms, like points, where it is the default. ⍰ position = “fill” works like stacking, but makes each set of stacked bars the same height. This makes it easier to compare proportions across groups.

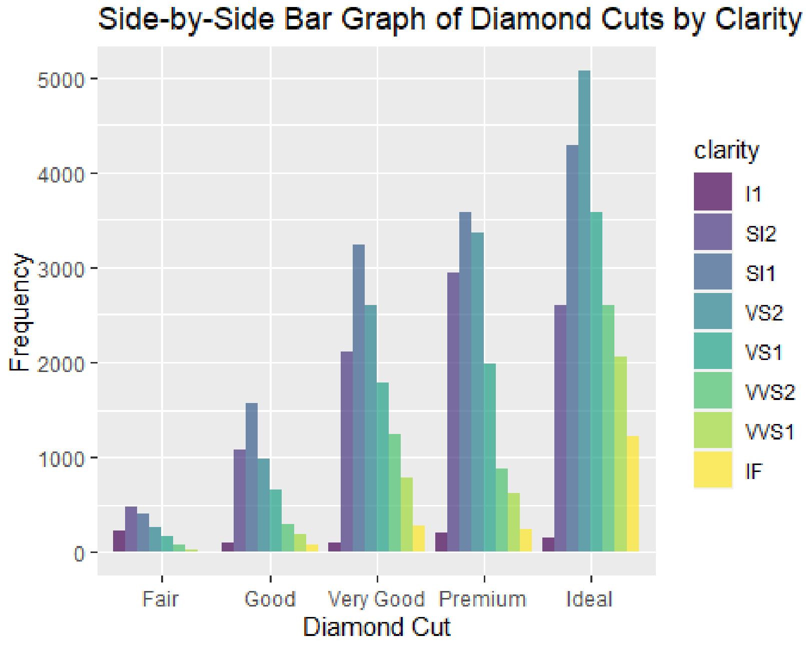
Using position = “fill” the bars fill the vertical space proportionally

```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = clarity), position = "fill") +
    labs(x = "Diamond Cut",
         y = "Proportion",
         title = "Proportional Bar Graph of Diamond Cuts by Clarity")
```



Position = “dodge” will get side-by-side bars

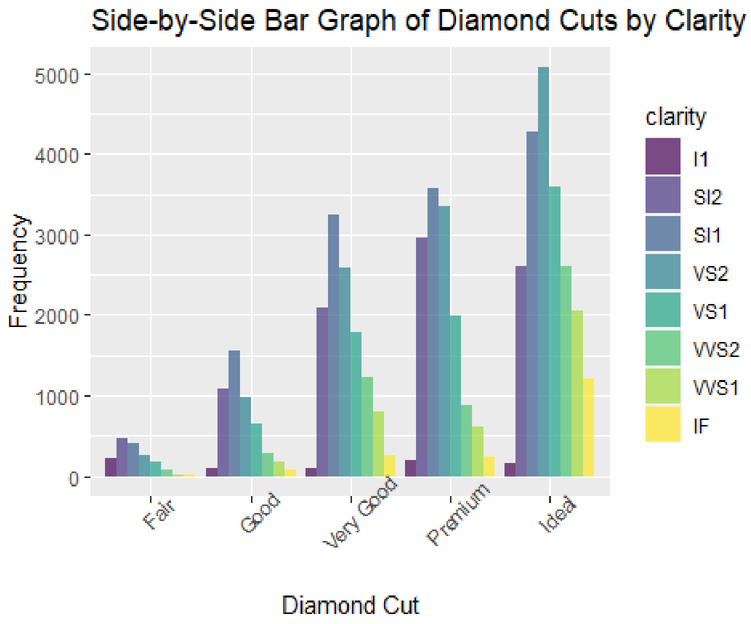
```
ggplot(data = diamonds, aes(x = cut, fill = clarity)) +
      geom_bar(alpha = .7, position = "dodge") +
    labs(x = "Diamond Cut",
         y = "Proportion",
         title = "Side-by-Side Bar Graph of Diamond Cuts by Clarity")
```



Change the angle of the x-axis labels

When x-axis labels are too long, they may overlap. You can change the text angle with axis.text.x = element_text(angle = 45))

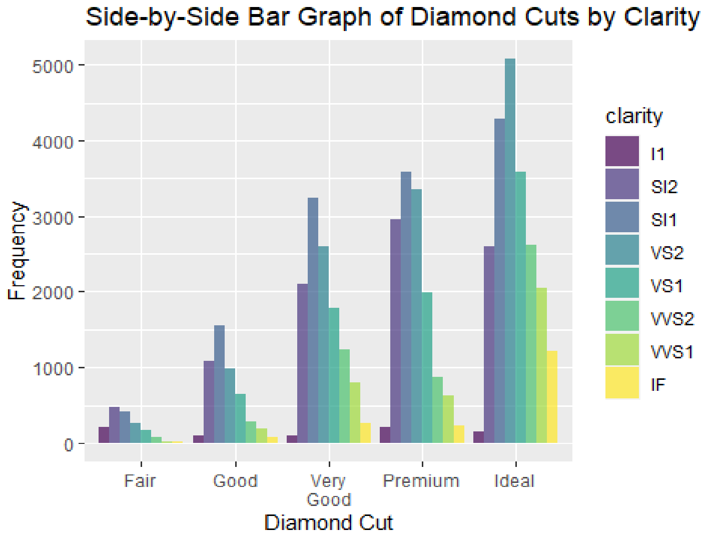
```
ggplot(data = diamonds, aes(x = cut, fill = clarity)) +
  geom_bar(alpha = .7, position = "dodge") +
    labs(x = "Diamond Cut",
         y = "Proportion",
        title = "Side-by-Side Bar Graph of Diamond Cuts by Clarity") +
    theme(axis.text.x = element_text(angle = 45))
```



Finally, make the x-axis labels fit in a narrow width

Here is another option for dealing with x-axis labels when they are long. You can use this function to break words into 2 lines.

```
ggplot(data = diamonds, aes(x = cut, fill = clarity)) +
  geom_bar(alpha = .7, position = "dodge") +
  labs(x = "Diamond Cut",
       y = "Proportion",
       title = "Side-by-Side Bar Graph of Diamond Cuts by Clarity") +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 5))
```



```
# notice "very good" will fit on two lines instead of one line
```

Homework Assignments Week 1

1. **(Ungraded)** Install R and R-Studio on your computer using the documents and videos provided. Sign up for a free Rpubs account. Start exploring R Studio by copying the code from the Markdown page into your own Markdown document.

2. **(Ungraded)** Reread these notes and try copying, pasting, and running the code provided above. Each week **YOU ARE RESPONSIBLE FOR LEARNING THE CODE PRESENTED IN THE NOTES**.

3. **(Worth 10 points)** Post your introduction video in the **discussion section on blackboard** (by Friday, Jan 31st). Respond to at least 3 classmates **by Sunday, Feb 2nd.**

4. **(Worth up to 10 points)** Explore the "reputable sites" listed in the notes above (NY Times, FiveThiryEight, BBC, HuffPost, Washington Post, Economist, Vox, etc.).

   a. Select one site and then select one relatively current visualization (you certainly may be interested in COVID-19 related visualizations).
   b. On a Word document, paste the image (must have color).
   c. On this same document, write a summary about this image.
   d. Describe what has been done well in this visualization.
   e. Describe what could be improved; it might show selection bias or any other type of distortion.
   f. Be sure to include a link/url to your data visualization (and be sure to put your name at the top)

Submit the graded assignment in the appropriate Blackboard Assignment Dropbox by **11:59 pm on Sunday, Feb 2nd.** We will present/discuss your submissions during the next class.