Econ5/Poli5D

David Arnold

2022-01-21

Contents

Introduction

This is an online book for ECON 5/ POLI 5D: Introduction to Social Data Analytics.

As data about individuals, organizations, and governments become increasingly available, social data analytics are transforming the way we think about the economy, politics and society. This course will teach skills necessary to navigate the world of social data. Each week we will cover an empirical application, often highlighting research by the faculty at UCSD. These applications will cover a wide variety of topics, including intergenerational mobility, voting, labor-market discrimination, among others.

To study these topics, we need to learn how to analyze data. In this class, we will learn about three softwares that are popular in social science: Excel, Stata, and R. While learning coding fundamentals, we will shed light on big social science questions.

[Other important course information here]

Chapter 1

Financial Incentives and Student Performance (Excel)

1.1 Introduction to Experiments

Did you know that countries with higher consumption of chocolate per capita tend to win more Nobel prizes?

In Figure 1.1, the number of Nobel laureates per 10 million of the population (vertical axis) is plotted against the chocolate consumption in kilograms per capita in a given year (horizontal axis). As is clear in this graph, countries with higher rates of chocolate consumption also tend to have higher rates of Nobel laureates is an example of a **correlation**. A correlation is any statistical is any relationship between two variables.

Correlations can be very interesting. In many cases they can lead to new research questions. But they don't often answer the questions social scientists are actually interested in. We are generally interested in understanding **causal relationships** between variables.

To understand causal relationships, let's again consider the relationship between chocolate consumption and Nobel prizes. Do you think that if the U.S. wanted to increase the number of Nobel laureates, they should invest in chocolate and include chocolate in lunch meals for students? Probably not! It's unlikely that increased chocolate consumption **causes** an individual to win a Nobel prize. In this case, it is more likely that this correlation in the data just arose due to chance. Therefore, we have uncovered a **spurious correlation**, or a correlation that arose due to chance. There are many interesting spurious correlations. If

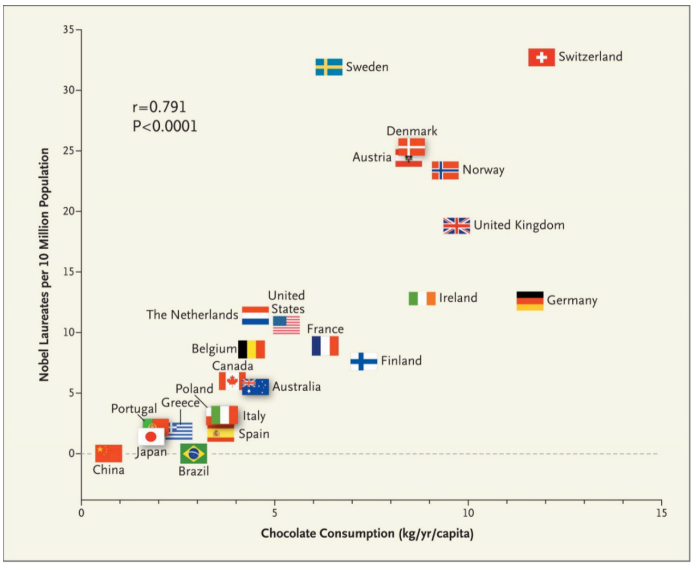


Figure 1.1: Chocolate Consumption and Nobel Prizes

you are interested in looking through some fun spurious correlations, just follow this link. For example, you probably would not guess that consumption of mozzarella cheese is correlated with civil engineering doctorates awarded, but see Figure 1.2 for this surprising correlation.

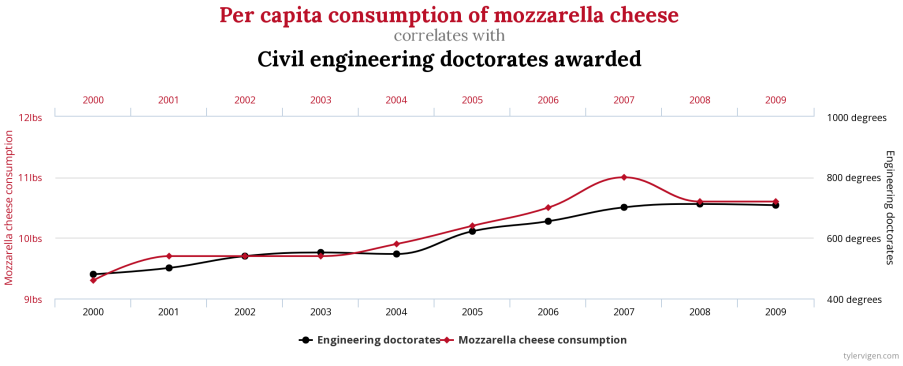


Figure 1.2: Example of a Spurious Correlation from tylervigen.com

If we need to decide whether a given policy is worth it or not, we need to understand how it causally impacts individual's outcomes. To see why, let's imagine a high school is trying to decide whether to invest in a new after-school program designed to help students with SAT prep. Currently, they do not have such a program on campus, but they know some of their students attend

programs offered outside the school. In order to try and understand the benefits of the program, the school decides to use data on past students to see if those who attended SAT prep programs outperformed students that did not.

The school finds those who attended SAT test prep programs score much higher on the SAT than those who did not. Does this mean the SAT test prep program works and the school should invest in one? Maybe, maybe not. There are two stories that could potentially rationalize this finding. Story 1 is that the programs are effective ways for students to learn the material they need to perform well on the SAT. That is, the program has a causal impact on SAT scores. Story 2 is that the programs are not actually effective. However, students that self-select into SAT programs also tend to be more motivated overall. They could, in theory, vary along many dimensions: academic ability, interest in college, parental background. It could be that any of these differences are actually driving the differences in SAT scores.

With only **observational data** it is difficult to tell which story is true. However, there is along-established way to identify causal relationship in medicine: the randomized control trial (RCT). To understand RCTs, imagine we are testing whether a given drug reduces blood pressure. To study the effectiveness we recruit a group of participants. Half of the participants are randomly placed into the **treatment** group and given the drug. The other half are placed into the **control** or **placebo** group and given a placebo drug. Now we can just compare outcomes for the two groups over time to understand the impact of the drug.

The key to a well-done experiment is **randomization**. If who receives the treatment is completely random, then on average, the two groups will be very similar. For example, in our test prep example, if we could randomly split students into two groups: test prep vs. no test prep, then we could evaluate the impacts of the program. Since students are no longer self-selecting into the program, there should be little difference in the two groups in terms of academic ability, college interest, parental background.

Not all questions in social science can be answered with an experiment. In certain cases it may be prohibitively expensive or unethical to run an experiment. However, if an experiment can be done, it is hard to find more convincing evidence. Experiments have been widely influential in many social sciences. For example, in the field of development economics, a large number of experiments have been undertaken in order to develop policies aimed at alleviating global povery. In 2019, Abhijit Banerjee, Esther Duflo, and Michael Kremer won the Nobel Prize in Economics "for their experimental approach to alleviating global poverty."

In this course, we will be studying a number of social science experiments. Our first experiment comes from Brownback and Sadoff (2020). The motivation for this experiment is a startling statistic. Only 40 percent of community college students earn a college degree within 6 years (Shapiro et al. 2017). Given these

poor outcomes, it is important to consider tools to improve performance. In Brownback and Sadoff (2020) the researchers explore one potential tool that has been previously unexplored: financial incentives to instructors. This is often referred to as a pay-for-performance model. Instructors are paid for how well their students perform. While prior experiments have studied this model in some settings, no prior work has explored how this would function in the community college setting.

Brownback and Sadoff (2020) run the following experiment at a community college in Indiana (Ivy Tech):

- Some instructors (the treatment group) received $50 for each student that passes an externally-administered test
- Other instructors (the control group) did not receive such an incentive.

The key in their design is that the instructors that are chosen to receive the incentives are randomly selected. Therefore, they are not, on average, better teachers than instructors in the control group. Therefore, if we observe any difference in student performance between the treatment and control, then it must be due to the financial incentives for instructors.

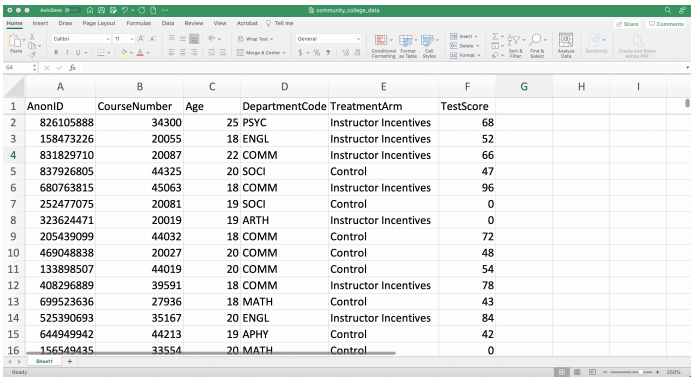
Summary: Social scientists are often interested in causal relationships. One way to study causal relationships is through an experiment. The key aspect in an experiment that allows us to study causal relationships is that **treatment is randomized**.

1.2 Data from Brownback and Sadoff (2020)

In many disciplines, there has been a push toward transparency and replication. Past work has shown that some very influential studies have failed to replicate (See replication crisis). A common way to provide transparency is to have authors make the data from their research publicly available. That means we will get to use the actual data from the experiments we are studying in this class.

Table 1.3 shows a selection of the data from Brownstone and Sadoff (2020). The actual data is much larger, but only a subset of the variables will be relevant for our analysis. This is referred to as a **data table**. Each **row** in a data table is an observation, while each **column** is a variable.

It is always important to understand the structure of your data before proceeding with any data analysis. One of the most important components of understanding the structure is to discern the **unit of observation**. The unit of observation is the level at which the data is reported. To discern it for yourself, just think about what each row of the dataset represent. Is each row an individual, a neighborhood, a country, a state? To practice, let's go through a few examples.



	A	B	C	D	E	F	G	H	I
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore			
2	826105888	34300	25	PSYC	Instructor Incentives	68			
3	158473226	20055	18	ENGL	Instructor Incentives	52			
4	831829710	20087	22	COMM	Instructor Incentives	66			
5	837926805	44325	20	SOCI	Control	47			
6	680763815	45063	18	COMM	Instructor Incentives	96			
7	252477075	20081	19	SOCI	Control	0			
8	323624471	20019	19	ARTH	Instructor Incentives	0			
9	205439099	44032	18	COMM	Control	72			
10	469048838	20027	20	COMM	Control	48			
11	133898507	44019	20	COMM	Control	54			
12	408296889	39591	18	COMM	Instructor Incentives	78			
13	699523636	27936	18	MATH	Control	43			
14	525390693	35167	20	ENGL	Instructor Incentives	84			
15	644949942	44213	19	APHY	Control	42			
16	156549435	33554	20	MATH	Control	0			

Figure 1.3: Data From Brownstone and Sadoff (2020)

What is the unit of observation in Table 1.4?

E	F
Student ID	GPA
1	3.4
2	3.8

Figure 1.4: Unit of Observation Example 1

To answer this question, we need to decide what a row represents. In this case, each row corresponds to a different student. Therefore the observation level is a student.

What is the unit of observation in Table 1.5?

Again, we need to decide what a row represents. In this case, each row corresponds to a different student in a different term. Therefore the observation level is a student-term.

Now let's think about what the unit of observation is in the Brownstone and Sadoff (2020) data. The variable **AnonID** is the identifier for a student. The variable **CourseNumber** is the identifier for a course. This dataset seems to be giving us the information for a student in a given course. Therefore the observation level is a student-course.

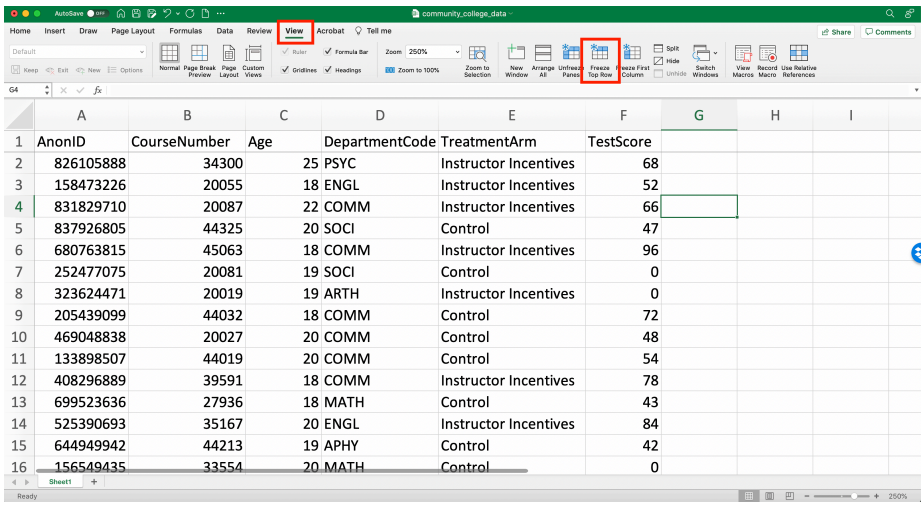
Before we get to analyzing the data, let's actually learn about our first feature in Excel. When browsing an excel file, it is often convenient to **freeze** the first row so that it is always visible. In our case, the first row holds variable names.

A	B	C
Student ID	Term	GPA
1	Fall	3
1	Winter	3.4
1	Spring	3.8
2	Fall	4
2	Winter	3.8
2	Spring	3.6

Figure 1.5: Unit of Observation Example 2

Therefore, if we freeze the first row we can scroll through the data without having to remember what each column corresponds to.

To freeze the first row go to the **View** and then click **Freeze Top Row** (See 1.6).



	A	B	C	D	E	F	G	H	I
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore			
2	826105888	34300	25	PSYC	Instructor Incentives	68			
3	158473226	20055	18	ENGL	Instructor Incentives	52			
4	831829710	20087	22	COMM	Instructor Incentives	66			
5	837926805	44325	20	SOCI	Control	47			
6	680763815	45063	18	COMM	Instructor Incentives	96			
7	252477075	20081	19	SOCI	Control	0			
8	323624471	20019	19	ARTH	Instructor Incentives	0			
9	205439099	44032	18	COMM	Control	72			
10	469048838	20027	20	COMM	Control	48			
11	133898507	44019	20	COMM	Control	54			
12	408296889	39591	18	COMM	Instructor Incentives	78			
13	699523636	27936	18	MATH	Control	43			
14	525390693	35167	20	ENGL	Instructor Incentives	84			
15	644949942	44213	19	APHY	Control	42			
16	156549435	33554	20	MATH	Control	0			

Figure 1.6: Freezing the First Row

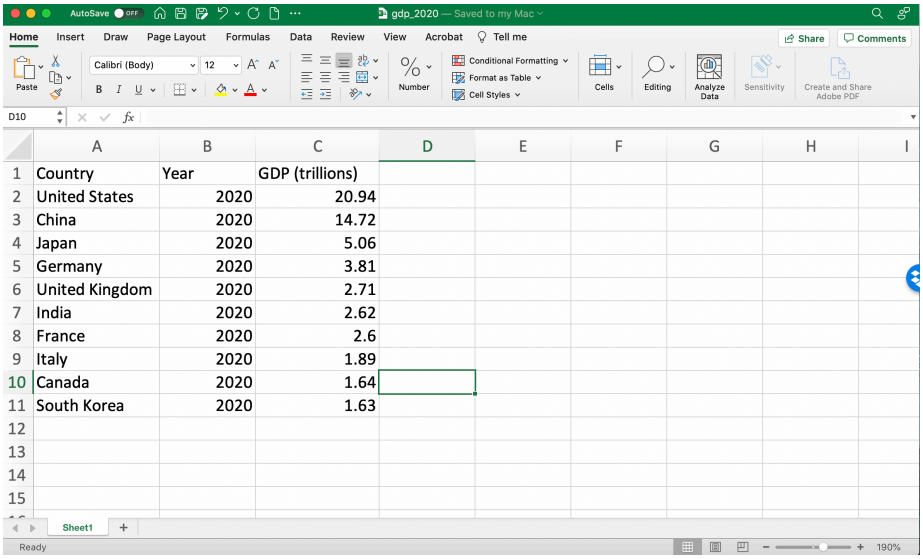
Let's talk a little bit about our goal with this dataset. Our goal is to determine if providing financial incentives to teachers impacts student performance. The variable **Treatment Arm** details whether a student has an instructor that received incentives or is in the control group. Therefore, we can see if test scores improve based on the Treatment Arm of the student. To perform this analysis, we need to learn how to use **functions** in Excel.

1.3 Statistical Functions

A function is something that takes in an input and produces an output. For example, you can think of taking the average as a function. The input is a list of numbers, and the output is the average of that list of numbers. When the inputs and outputs of a function are numbers (as it is in this example), then it is a **statistical function**.

In Excel, statistical functions are extremely important to understand. If you are doing research, it is common to start out with a list of summary statistics. If you are in business, it might be important to know some summary statistics about your products, sales, costs, etc. To create summary statistics these summary statistics, we use statistical functions.

In order to illustrate the use of statistical functions, let's consider the following dataset that has information on Gross Domestic Product (GDP) for ten countries



	A	B	C	D	E	F	G	H	I
1	Country	Year	GDP (trillions)						
2	United States	2020	20.94						
3	China	2020	14.72						
4	Japan	2020	5.06						
5	Germany	2020	3.81						
6	United Kingdom	2020	2.71						
7	India	2020	2.62						
8	France	2020	2.6						
9	Italy	2020	1.89						
10	Canada	2020	1.64						
11	South Korea	2020	1.63						
12									
13									
14									
15									

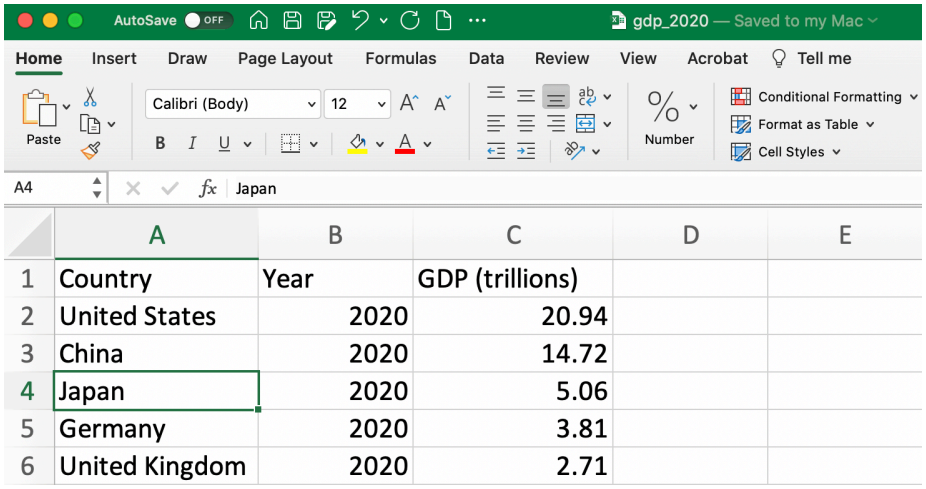
Figure 1.7: GDP in 2020

Let's introduce our first statistical function: the `AVERAGE` function. This function will take in a list of numbers and output the average of all those numbers. There are many statistical functions in Excel, including:

- `COUNT` – counts the number of numbers
- `SUM` – adds up the numbers
- `AVERAGE` – averages the numbers
- `MEDIAN` – retrieves the median of all numbers

- `MAX`, `MIN` – retrieves the maximum and minimum of all numbers respectively
- `MODE` – retrieves the mode of all numbers

In order to make use of statistical functions, we first need to learn how to reference cells. In Excel, every cell is identified by a column letter and a row number. For example, in 1.8 I have clicked on the cell that holds the word "Japan". This entry is in column A of row 4. Therefore, this is cell A4.



	A	B	C	D	E
1	Country	Year	GDP (trillions)		
2	United States	2020	20.94		
3	China	2020	14.72		
4	Japan	2020	5.06		
5	Germany	2020	3.81		
6	United Kingdom	2020	2.71		

Figure 1.8: How to Reference a Cell

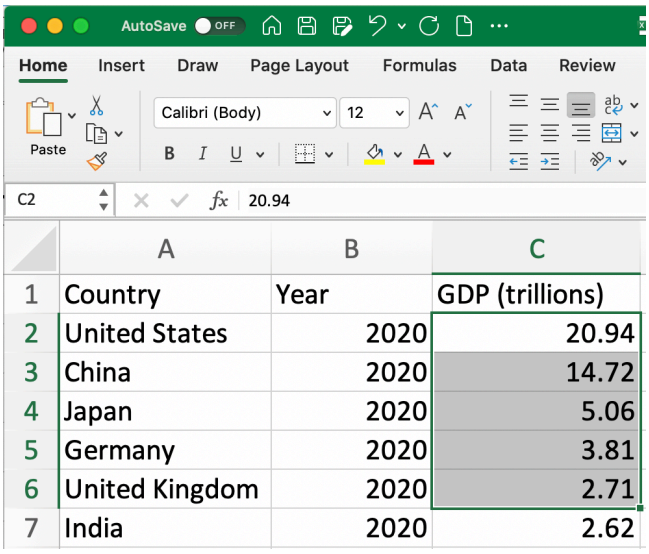
You can also reference a range of cells (this will be useful when taking averages). For example, to reference rows 2-6 of column C we could type `C2:C6` (See 1.9). Whenever you read a colon in Excel, you should read it as "through". Therefore, the text `C2:C6` can be read as "cells C2 through C6".

Sometimes, it is helpful to reference an entire column. For example, later we will open an Excel spreadsheet that has thousands of students. In some cases, it will be helpful to reference an entire column when using statistical functions. For example, if we are interested in the average of a variable in column C, we can type `C:C` to reference the entire column.

Now that we understand how to reference cells, we can begin to apply functions. We will begin with the `AVERAGE` functions which can compute an average of a list of numbers. In our spreadsheet, we are interested in computing the average GDP across these ten countries. To do so, we can first click in an unpopulated cell and type:

```
=AVERAGE(C:C)
```

Note the `=` sign in the text above. This tells Excel to evaluate the function. If you omit the `=` you will find that the cell is just populated with the text



	A	B	C
1	Country	Year	GDP (trillions)
2	United States	2020	20.94
3	China	2020	14.72
4	Japan	2020	5.06
5	Germany	2020	3.81
6	United Kingdom	2020	2.71
7	India	2020	2.62

Figure 1.9: Referencing a Range of Cells

`AVERAGE(C:C)` and not the average value of column C.

Figure 1.10 depicts an example where we have set up a Summary Statistics table for this dataset. In cell F3, we have written the function that calculates the average GDP across these ten countries.

As soon as we present enter, the function will be evaluated and the average will appear in cell F3 (see Figure 1.11).

The rest of the summary statistics are completely analogous, replacing `AVERAGE` with the relevant function. See Figure 1.12 for the output of the rest of the Summary Statistics table.

1.4 Logical Functions

Another type of function in Excel is called a **logical function**. Before learning about **logical functions** we need to learn more about logical statements in general. A logical statement is a statement that is either true or false. For example, all of the following statements are either true or false.

- "The student passed the class"
- "The participant is in the treatment group"
- "The temperature is below freezing"

These statements are all either true or false. This type of binary logic is incredibly important in certain areas of mathematics as well as the development of

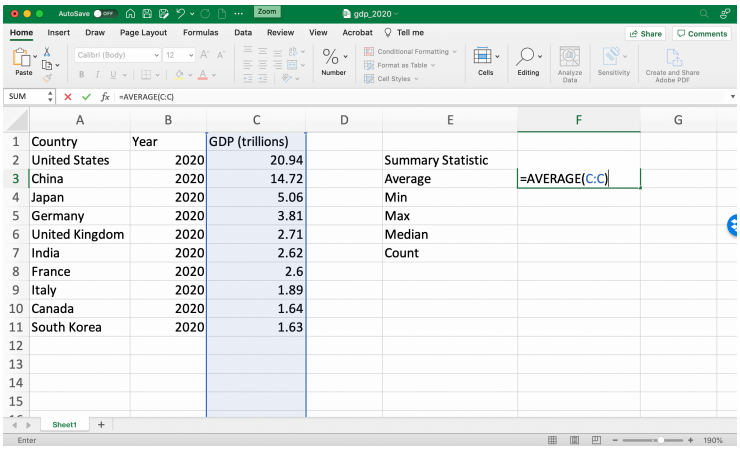


Figure 1.10: Computing Average GDP in 2020 (Before Pressing Enter)

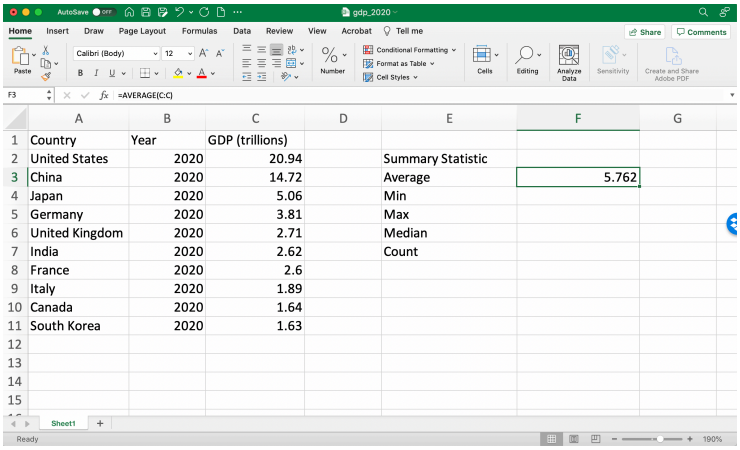
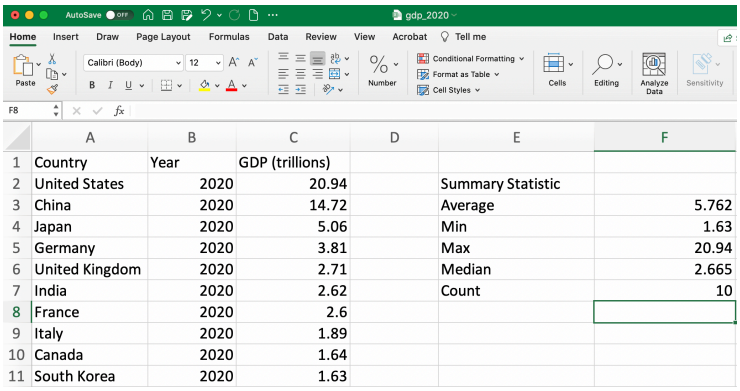


Figure 1.11: Computing Average GDP in 2020 (After Pressing Enter)



	A	B	C	D	E	F
1	Country	Year	GDP (trillions)			
2	United States	2020	20.94		Summary Statistic	
3	China	2020	14.72		Average	5.762
4	Japan	2020	5.06		Min	1.63
5	Germany	2020	3.81		Max	20.94
6	United Kingdom	2020	2.71		Median	2.665
7	India	2020	2.62		Count	10
8	France	2020	2.6			
9	Italy	2020	1.89			
10	Canada	2020	1.64			
11	South Korea	2020	1.63			

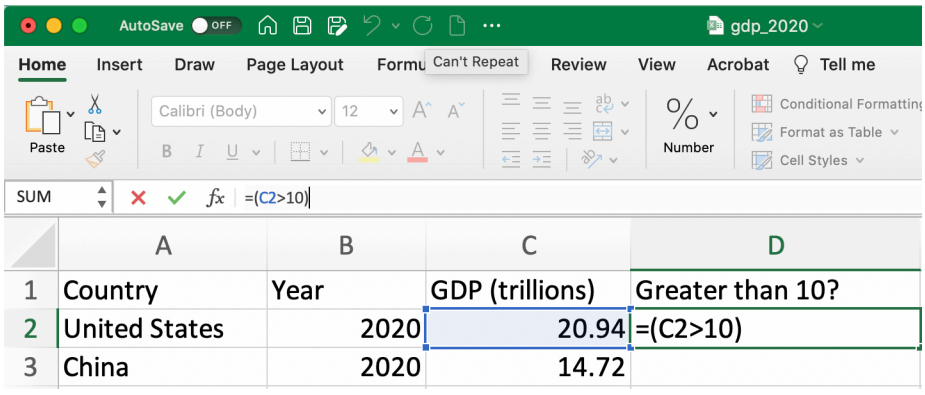
Figure 1.12: Summary Statistics Table

computers. An entire branch of mathematics, termed **Boolean algebra**, deals with variables that take on binary values.

In Excel, we can type logical statements and Excel will evaluate whether the statement is True or False. The key is you need to type the statement in the correct format so that Excel understands what you are trying to ask.

For example, in our dataset on GDP in 2020, one example of a logical statement is: "GDP is greater than 10 trillion dollars." But how do we type the statement "GDP is greater than 10 trillion dollars."

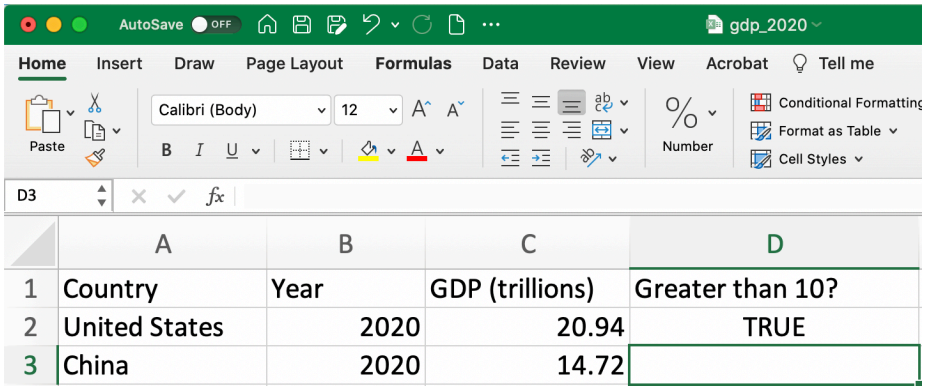
Figure 1.13 depicts how we would type this statement for the first observation in the dataset (the U.S.). To evaluate whether GDP is greater than 10 trillion for the U.S. we type `=(C2>10)`. The equals sign tells Excel to evaluate the logical statement that follows. The logical statement that follows `C2>10` is simply asking whether the value in cell C2 (i.e. U.S.'s GDP) is greater than 10.



	A	B	C	D
1	Country	Year	GDP (trillions)	Greater than 10?
2	United States	2020	20.94	=(C2>10)
3	China	2020	14.72	

Figure 1.13: Is GDP Greater than 10 in the U.S. Part 1

Once we press enter, we will get the result. In this case, we can see GDP in the U.S. is indeed greater than 10 trillion dollars. Therefore, when the function in cell D2 is evaluated it returns the answer `TRUE` as can be seen in Figure 1.14.



	A	B	C	D
1	Country	Year	GDP (trillions)	Greater than 10?
2	United States	2020	20.94	TRUE
3	China	2020	14.72	

Figure 1.14: Is GDP Greater than 10 in the U.S. Part 2

Once we have filled in the formula for one cell, we can also apply it to the remaining observations. In our example, this would mean testing whether each country's GDP is more than 10, not just the U.S. There are a couple of ways to do this efficiently in Excel.

One way to fill in the formula is to click the bottom right corner of cell D2 and then drag down. This allows you to apply to as many observations as you would like. However, we often want to apply the logical statement to all observations. Therefore, continuously dragging will be inefficient if there are thousands of observations in the data. To quickly apply a formula to all cells in a column, we can simply double click the bottom right corner of a cell. In our example, this would mean double clicking the bottom right corner of the D2 cell.

Oftentimes, instead of returning "TRUE" or "FALSE", it is convenient to have Excel return the value in another form. For example, a common way to represent "TRUE" is with the number 1 and "FALSE" with the number 0. To accomplish this in Excel we will use the `IF` function.

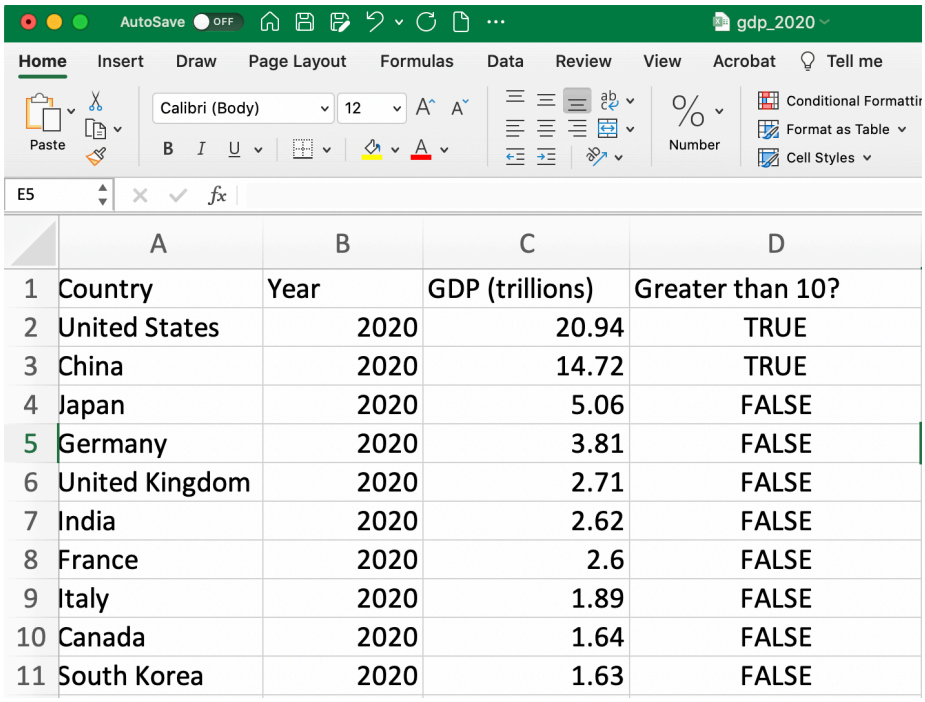
The basic format for the `IF` function is:

```
=IF(logical test,value if TRUE, value if FALSE)
```

For example, in our dataset, if we want to return the value of 1 if GDP is greater than 10 and 0 if otherwise, we can type:

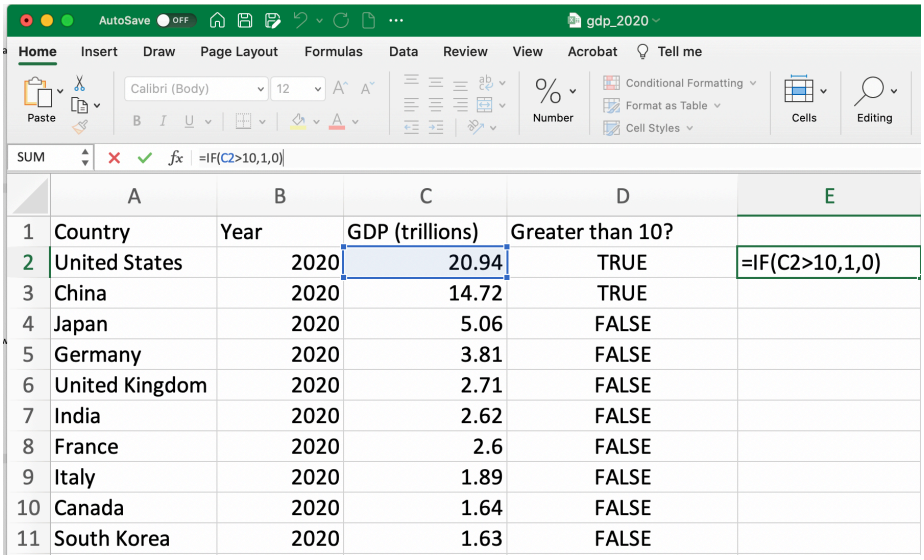
```
=IF(C2>10,1,0)
```

Once we press enter, the function is evaluated. To fill in the formula for the rest of the observations you can follow the same steps as above.



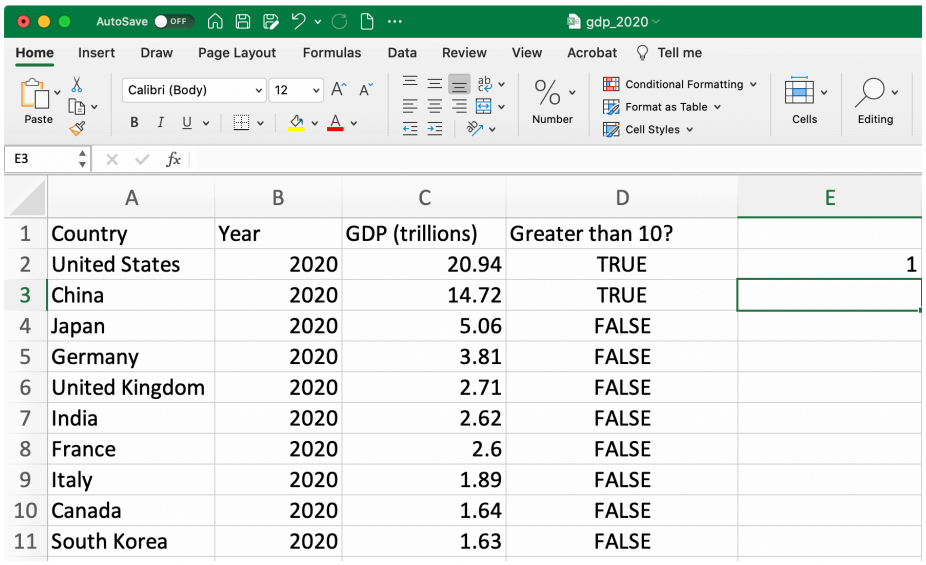
	A	B	C	D
1	Country	Year	GDP (trillions)	Greater than 10?
2	United States	2020	20.94	TRUE
3	China	2020	14.72	TRUE
4	Japan	2020	5.06	FALSE
5	Germany	2020	3.81	FALSE
6	United Kingdom	2020	2.71	FALSE
7	India	2020	2.62	FALSE
8	France	2020	2.6	FALSE
9	Italy	2020	1.89	FALSE
10	Canada	2020	1.64	FALSE
11	South Korea	2020	1.63	FALSE

Figure 1.15: Filling in the Formula



	A	B	C	D	E
1	Country	Year	GDP (trillions)	Greater than 10?	
2	United States	2020	20.94	TRUE	=IF(C2>10,1,0)
3	China	2020	14.72	TRUE	
4	Japan	2020	5.06	FALSE	
5	Germany	2020	3.81	FALSE	
6	United Kingdom	2020	2.71	FALSE	
7	India	2020	2.62	FALSE	
8	France	2020	2.6	FALSE	
9	Italy	2020	1.89	FALSE	
10	Canada	2020	1.64	FALSE	
11	South Korea	2020	1.63	FALSE	

Figure 1.16: Using the IF function (Before Pressing Enter)



	A	B	C	D	E
1	Country	Year	GDP (trillions)	Greater than 10?	
2	United States	2020	20.94	TRUE	1
3	China	2020	14.72	TRUE	
4	Japan	2020	5.06	FALSE	
5	Germany	2020	3.81	FALSE	
6	United Kingdom	2020	2.71	FALSE	
7	India	2020	2.62	FALSE	
8	France	2020	2.6	FALSE	
9	Italy	2020	1.89	FALSE	
10	Canada	2020	1.64	FALSE	
11	South Korea	2020	1.63	FALSE	

Figure 1.17: Using the IF function (After Pressing Enter)

1.5 Summary Statistics

Now let's return to our data from Brownback and Sadoff (2020). Our goal is to create a table of basic summary statistics. However, a key variable in the Brownback and Sadoff (2020) experiment is whether a given student passed the test or not. This financial compensation of an instructor (in the treatment group) directly depends on how many students pass the test. This currently does not appear in our dataset, so we will have to create it.

To create this variable we need to take three steps:

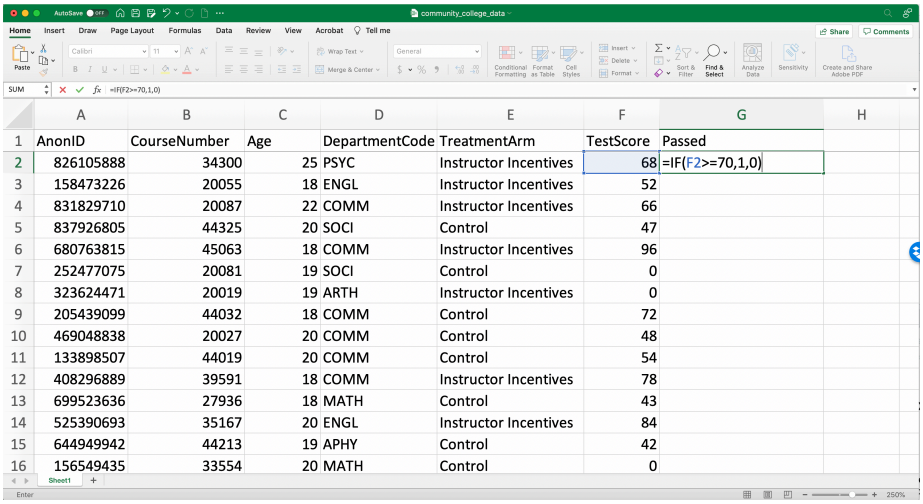
- **First Step:** Title the column with the variable name "Passed"
- **Second Step:** Use the `IF` function to fill in this variable for the first observation
- **Third Step:** Double click bottom right of cell in step 2 to fill in the variable for all observations

Figure 1.18 presents steps 1 and 2 in Excel.

We simply can double click the bottom right of G2 to fill in the rest of the observations (as seen in Figure 1.19).

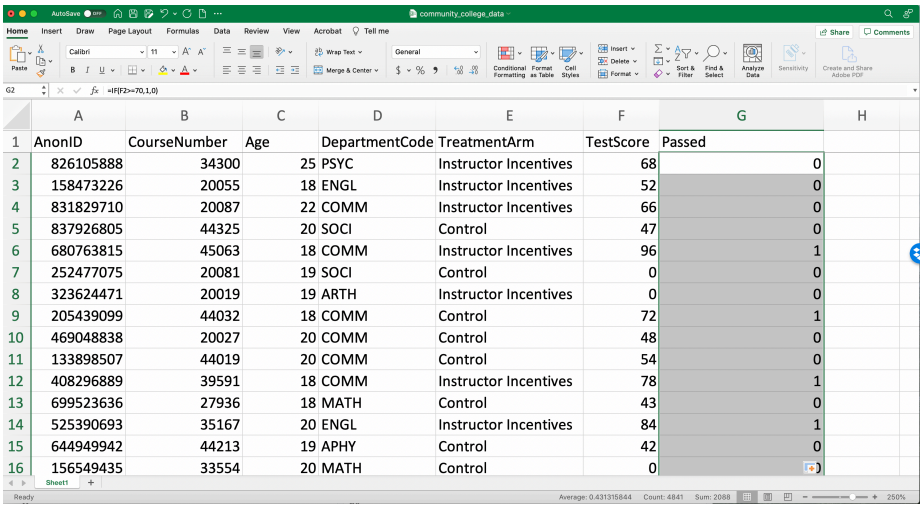
Now that we've generated the variables we need let's fill in some summary statistics displayed on the right side of the datasheet in Figure 1.20

To compute average age in the dataset we just need to use the `AVERAGE` function and reference column C.



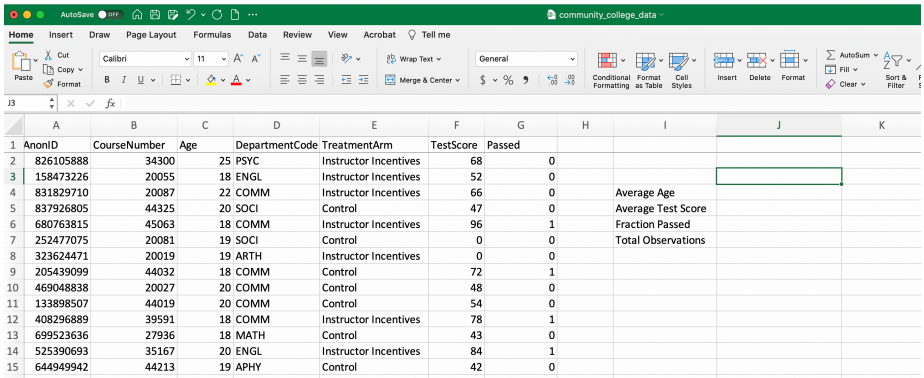
	A	B	C	D	E	F	G	H
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed	
2	826105888	34300	25	PSYC	Instructor Incentives	68	=IF(F2>=70,1,0)	
3	158473226	20055	18	ENGL	Instructor Incentives	52		
4	831829710	20087	22	COMM	Instructor Incentives	66		
5	837926805	44325	20	SOCI	Control	47		
6	680763815	45063	18	COMM	Instructor Incentives	96		
7	252477075	20081	19	SOCI	Control	0		
8	323624471	20019	19	ARTH	Instructor Incentives	0		
9	205439099	44032	18	COMM	Control	72		
10	469048838	20027	20	COMM	Control	48		
11	133898507	44019	20	COMM	Control	54		
12	408296889	39591	18	COMM	Instructor Incentives	78		
13	699523636	27936	18	MATH	Control	43		
14	525390693	35167	20	ENGL	Instructor Incentives	84		
15	644949942	44213	19	APHY	Control	42		
16	156549435	33554	20	MATH	Control	0		

Figure 1.18: Generating a Passed Variable (Steps 1 and 2)



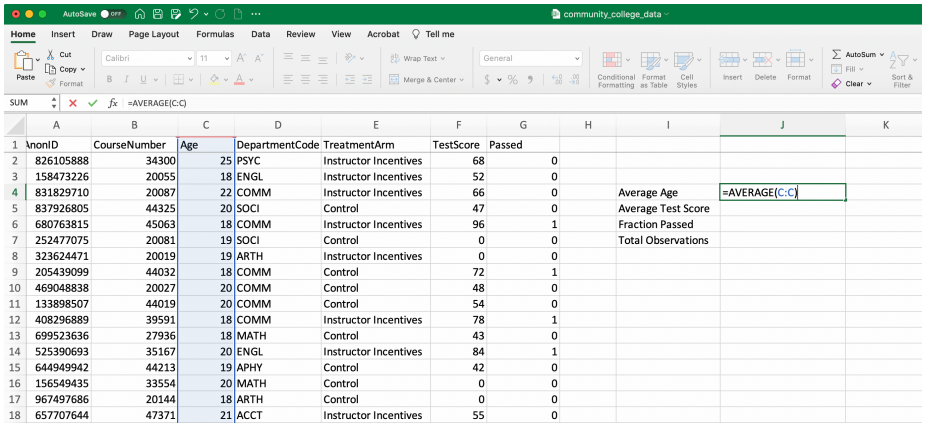
	A	B	C	D	E	F	G	H
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed	
2	826105888	34300	25	PSYC	Instructor Incentives	68	0	
3	158473226	20055	18	ENGL	Instructor Incentives	52	0	
4	831829710	20087	22	COMM	Instructor Incentives	66	0	
5	837926805	44325	20	SOCI	Control	47	0	
6	680763815	45063	18	COMM	Instructor Incentives	96	1	
7	252477075	20081	19	SOCI	Control	0	0	
8	323624471	20019	19	ARTH	Instructor Incentives	0	0	
9	205439099	44032	18	COMM	Control	72	1	
10	469048838	20027	20	COMM	Control	48	0	
11	133898507	44019	20	COMM	Control	54	0	
12	408296889	39591	18	COMM	Instructor Incentives	78	1	
13	699523636	27936	18	MATH	Control	43	0	
14	525390693	35167	20	ENGL	Instructor Incentives	84	1	
15	644949942	44213	19	APHY	Control	42	0	
16	156549435	33554	20	MATH	Control	0		

Figure 1.19: Generating a Passed Variable (Step 3)



	A	B	C	D	E	F	G	H	I	J	K
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed				
2	826105888	34300	25	PSYC	Instructor Incentives	68	0				
3	158473226	20055	18	ENGL	Instructor Incentives	52	0				
4	831829710	20087	22	COMM	Instructor Incentives	66	0		Average Age		
5	837926805	44325	20	SOCI	Control	47	0		Average Test Score		
6	680763815	45063	18	COMM	Instructor Incentives	96	1		Fraction Passed		
7	252477075	20081	19	SOCI	Control	0	0		Total Observations		
8	323624471	20019	19	ARTH	Instructor Incentives	0	0				
9	205439099	44032	18	COMM	Control	72	1				
10	469048838	20027	20	COMM	Control	48	0				
11	133898507	44019	20	COMM	Control	54	0				
12	408296889	39591	18	COMM	Instructor Incentives	78	1				
13	699523636	27936	18	MATH	Control	43	0				
14	525390693	35167	20	ENGL	Instructor Incentives	84	1				
15	644949942	44213	19	APHY	Control	42	0				

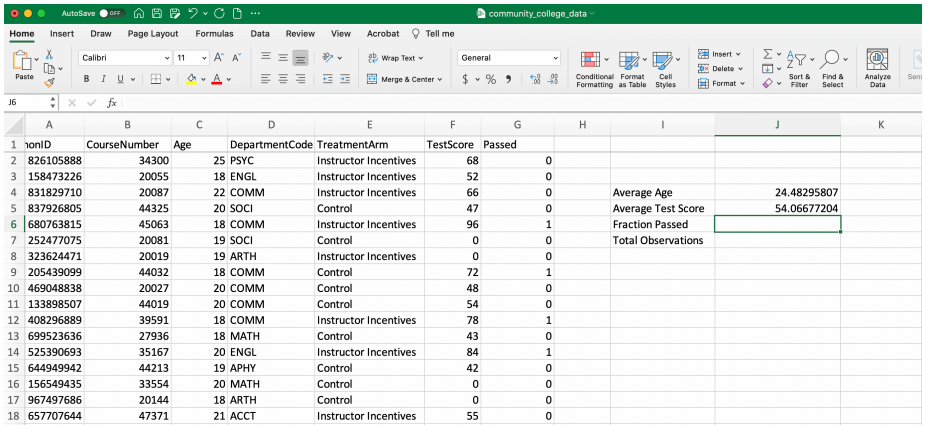
Figure 1.20: Blank Summary Statistics Table



	A	B	C	D	E	F	G	H	I	J	K
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed				
2	826105888	34300	25	PSYC	Instructor Incentives	68	0				
3	158473226	20055	18	ENGL	Instructor Incentives	52	0				
4	831829710	20087	22	COMM	Instructor Incentives	66	0		Average Age	=AVERAGE(C:C)	
5	837926805	44325	20	SOCI	Control	47	0		Average Test Score		
6	680763815	45063	18	COMM	Instructor Incentives	96	1		Fraction Passed		
7	252477075	20081	19	SOCI	Control	0	0		Total Observations		
8	323624471	20019	19	ARTH	Instructor Incentives	0	0				
9	205439099	44032	18	COMM	Control	72	1				
10	469048838	20027	20	COMM	Control	48	0				
11	133898507	44019	20	COMM	Control	54	0				
12	408296889	39591	18	COMM	Instructor Incentives	78	1				
13	699523636	27936	18	MATH	Control	43	0				
14	525390693	35167	20	ENGL	Instructor Incentives	84	1				
15	644949942	44213	19	APHY	Control	42	0				
16	156549435	33554	20	MATH	Control	0	0				
17	967497686	20144	18	ARTH	Control	0	0				
18	657707644	47371	21	ACCT	Instructor Incentives	55	0				

Figure 1.21: Average Age

Average test score is completely analogous to age, but we now reference column F.



	A	B	C	D	E	F	G	H	I	J	K
1	nonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed				
2	826105888	34300	25	PSYC	Instructor Incentives	68	0				
3	158473226	20055	18	ENGL	Instructor Incentives	52	0				
4	831829710	20087	22	COMM	Instructor Incentives	66	0		Average Age	24.48295807	
5	837926805	44325	20	SOCI	Control	47	0		Average Test Score	54.06677204	
6	680763815	45063	18	COMM	Instructor Incentives	96	1		Fraction Passed		
7	252477075	20081	19	SOCI	Control	0	0		Total Observations		
8	323624471	20019	19	ARTH	Instructor Incentives	0	0				
9	205439099	44032	18	COMM	Control	72	1				
10	469048838	20027	20	COMM	Control	48	0				
11	133898507	44019	20	COMM	Control	54	0				
12	408296889	39591	18	COMM	Instructor Incentives	78	1				
13	699523636	27936	18	MATH	Control	43	0				
14	525390693	35167	20	ENGL	Instructor Incentives	84	1				
15	644949942	44213	19	APHY	Control	42	0				
16	156549435	33554	20	MATH	Control	0	0				
17	967497686	20144	18	ARTH	Control	0	0				
18	657707644	47371	21	ACCT	Instructor Incentives	55	0				

Figure 1.22: Average Test Score

In order to compute fraction that passed the test, we are going to make a small digression, one that will clarify one reason it was convenient to store the variable Passed with a 1 if the individual passed and a zero otherwise. Because we have stored the variable in this way, we can simply take the average value of Passed and this will be equal to the fraction that passed the test.

To see why this is true, let's first write out the equation for the average of a generic variable X

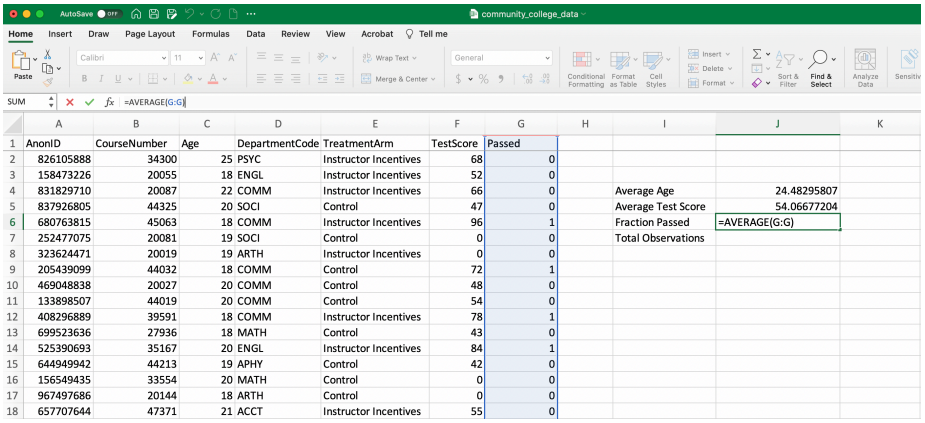
$$\bar{X} = \frac{X_1 + ... + X_N}{N}$$

Where N is the number of observations, X_1 is the first observation and X_N is the N^{th} observation. If X is a binary variable that is equal to 1 or zero, then the average will be the fraction of individual's with a 1

$$\bar{X} = \frac{X_1 + ... + X_N}{N} = \frac{\text{Number of observations } =1}{\text{Total Observations}}$$

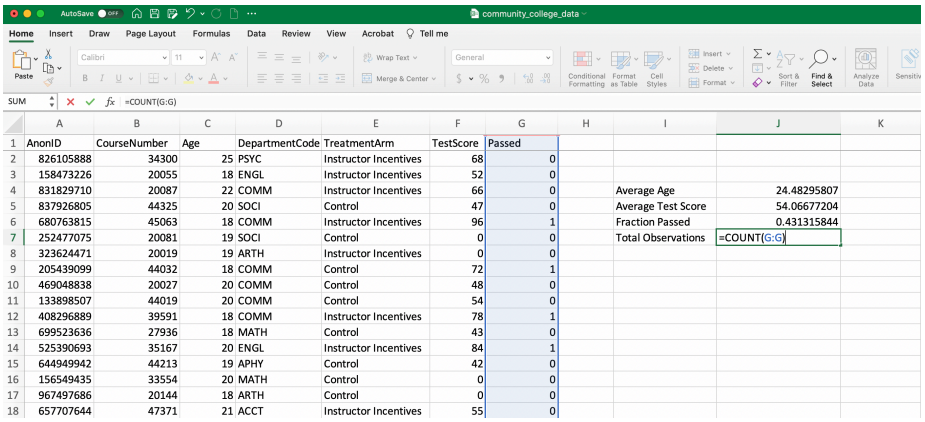
Since our variable "Passed" is equal to 1 if passed and zero otherwise, the fraction who passed is the average of our "Passed" variable. Therefore, in Excel we simply need to compute the average of the G column

Finally, to finish our summary statistics table, we need to fill in the total number of observations. To do this, we can use the `COUNT` function, which will count the total number of numbers in a column. While Figure 1.24 uses Column G to compute total observations, this could be replaced with any column that has non-missing numeric data.



	A	B	C	D	E	F	G	H	I	J	K
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed				
2	826105888	34300	25	PSYC	Instructor Incentives	68	0				
3	158473226	20055	18	ENGL	Instructor Incentives	52	0				
4	831829710	20087	22	COMM	Instructor Incentives	66	0		Average Age	24.48295807	
5	837926805	44325	20	SOCI	Control	47	0		Average Test Score	54.06677204	
6	680763815	45063	18	COMM	Instructor Incentives	96	1		Fraction Passed	=AVERAGE(G:G)	
7	252477075	20081	19	SOCI	Control	0	0		Total Observations		
8	323624471	20019	19	ARTH	Instructor Incentives	0	0				
9	205439099	44032	18	COMM	Control	72	1				
10	469048838	20027	20	COMM	Control	48	0				
11	133898507	44019	20	COMM	Control	54	0				
12	408296889	39591	18	COMM	Instructor Incentives	78	1				
13	699523636	27936	18	MATH	Control	43	0				
14	525390693	35167	20	ENGL	Instructor Incentives	84	1				
15	644949942	44213	19	APHY	Control	42	0				
16	156549435	33554	20	MATH	Control	0	0				
17	967497686	20144	18	ARTH	Control	0	0				
18	657707644	47371	21	ACCT	Instructor Incentives	55	0				

Figure 1.23: Computing the Fraction Passed



	A	B	C	D	E	F	G	H	I	J	K
1	AnonID	CourseNumber	Age	DepartmentCode	TreatmentArm	TestScore	Passed				
2	826105888	34300	25	PSYC	Instructor Incentives	68	0				
3	158473226	20055	18	ENGL	Instructor Incentives	52	0				
4	831829710	20087	22	COMM	Instructor Incentives	66	0		Average Age	24.48295807	
5	837926805	44325	20	SOCI	Control	47	0		Average Test Score	54.06677204	
6	680763815	45063	18	COMM	Instructor Incentives	96	1		Fraction Passed	0.431315844	
7	252477075	20081	19	SOCI	Control	0	0		Total Observations	=COUNT(G:G)	
8	323624471	20019	19	ARTH	Instructor Incentives	0	0				
9	205439099	44032	18	COMM	Control	72	1				
10	469048838	20027	20	COMM	Control	48	0				
11	133898507	44019	20	COMM	Control	54	0				
12	408296889	39591	18	COMM	Instructor Incentives	78	1				
13	699523636	27936	18	MATH	Control	43	0				
14	525390693	35167	20	ENGL	Instructor Incentives	84	1				
15	644949942	44213	19	APHY	Control	42	0				
16	156549435	33554	20	MATH	Control	0	0				
17	967497686	20144	18	ARTH	Control	0	0				
18	657707644	47371	21	ACCT	Instructor Incentives	55	0				

Figure 1.24: Computing Total Observations

1.6 Pivot Tables

So far, we have learned to retrieve summary statistics using functions. This works very well for retrieving summary statistics for the entire dataset. Sometimes however we want more complicated summary statistics. For example, in our empirical application, we might be interested in passing rates for different departments. While it is possible to retrieve these using functions, an easier way is through a **pivot table**. The name "pivot" comes from "pivoting" the data into a way the provides useful information.

To understand what we are trying to accomplish with our pivot table, let's start by looking at our end goal in Figure 1.25. In the Pivot Table, each **row** is a separate department, with the fraction passed displayed.

Row Labels	Average of Passed
ACCT	0.207446809
APHY	0.301923077
ARTH	0.317241379
BIOL	0.438202247
BOAT	0.570093458
BUSN	0.381231672
COMM	0.355191257
CRIM	0.546666667
ENGL	0.447162427
HLHS	0.636678201
MATH	0.4125
NRSG	0.979591837
PSYC	0.464367816
SCIN	0.392857143
SDEV	0.690909091
SOCI	0.492307692
(blank)	
Grand Total	**0.431315844**

Figure 1.25: Pivot Table: Pass Rates by Department

Now let's go through how to construct Figure 1.25. To insert a pivot table, go to the Insert tab and click "Insert Pivot Table", as seen in Figure 1.26

Next, you need to select a table or range. You should select all of the variables that you want to be in your pivot table. In our case, we need to have both the department and whether an individual passed. To select a range, we can click on a column (for example D), hold down, and then drag to another column (for example G). Since we began at D and dragged to G, the variables `DepartmentCode`, `TreatmentArm`, `TestScore`, and `Passed` will all be variables that we can include in the pivot table.

You can also choose where you would like your pivot table to appear. By default,

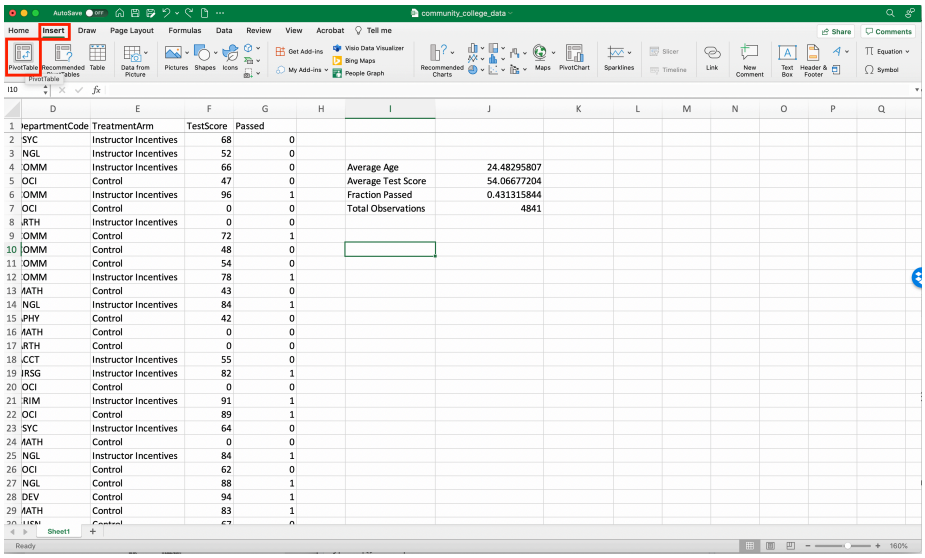


Figure 1.26: Inserting a Pivot Table

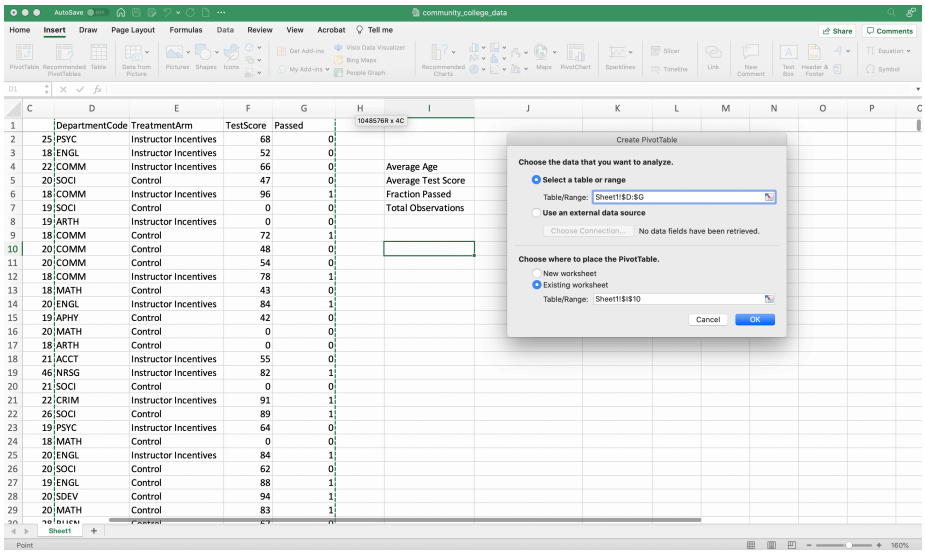


Figure 1.27: Selecting a Range for the Pivot Table

the table will be placed in whichever cell you clicked last (in my case this is cell `I10`). Therefore, before you insert a pivot table, you should click the cell where you would like it to appear.

As can be seen in Figure 1.28, Excel has drawn a blank Pivot table. On the right we can see the new panel "PivotTable Fields" which we will use to build our Pivot table

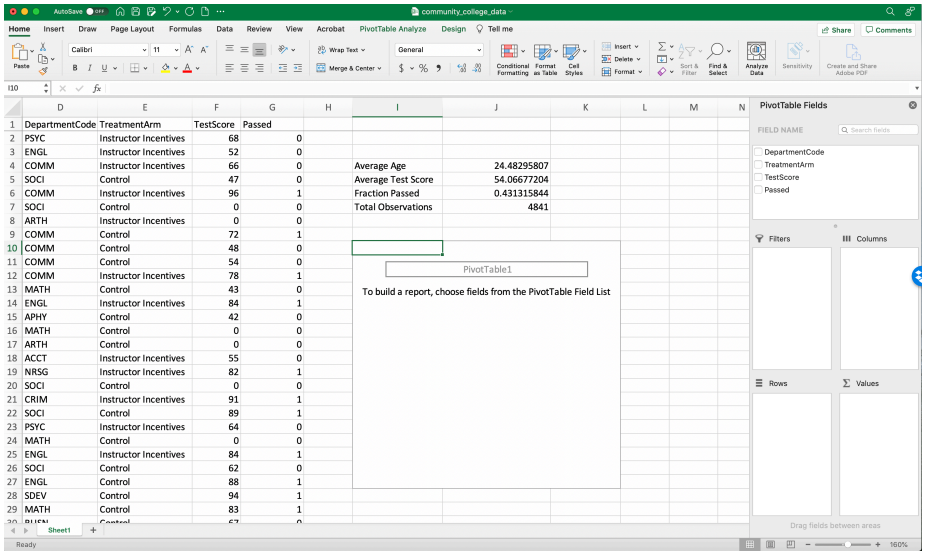


Figure 1.28: A Blank Pivot Table

We want each row in our pivot table to be a Department Code. So click the box next to `DepartmentCode` to put it into the pivot table.

By default (see Figure 1.29), Excel has placed this variable into the Values panel, but we want it to be in the Rows panel. We can move variables between these windows by clicking and dragging. In this example we want to move `DepartmentCode` to the Rows panel.

In the Values panel, we want pass rates. We will compute these pass rates from the variable "Passed"

By default Excel is displaying the "SUM" by department (i.e. number of individuals that passed, not the fraction that passed). To change the displayed statistic click the "i" next to the "Values" label and then select "Average".

Now we've reached our goal! A table that shows average pass rates by department (Figure 1.25).

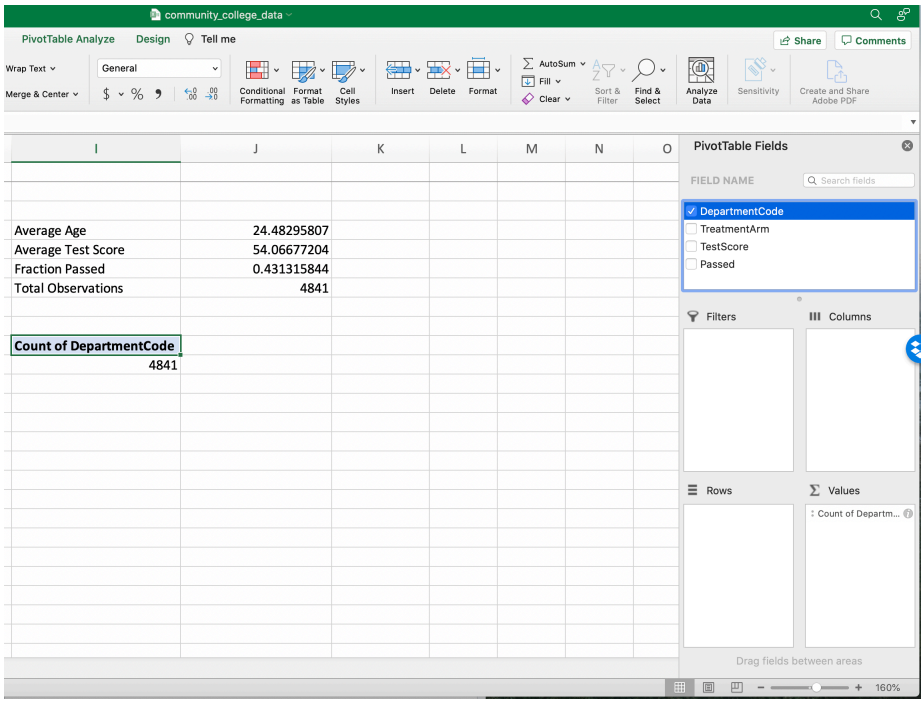


Figure 1.29: Adding DepartmentCode to the Pivot Table

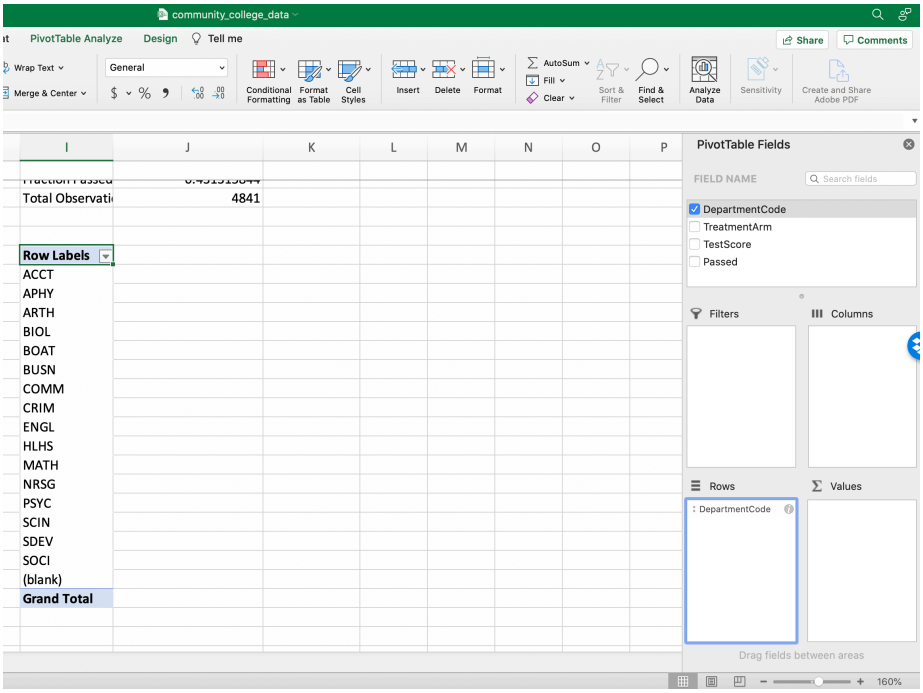


Figure 1.30: Adding DepartmentCode to the Pivot Table

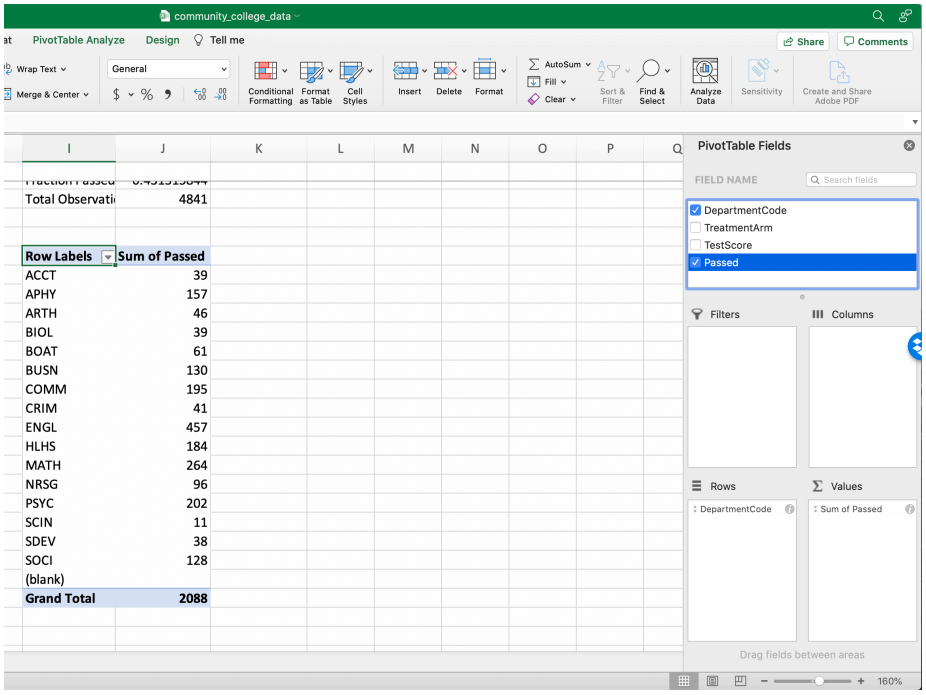


Figure 1.31: Adding Passed to the Pivot Table

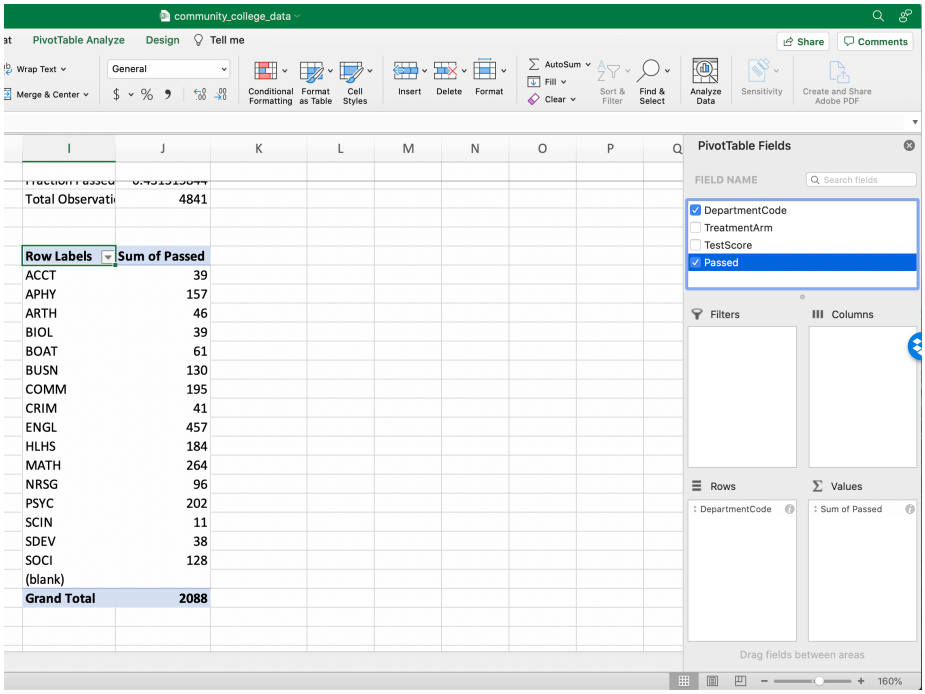


Figure 1.32: Average Pass Rates

1.7 Balance Tables

The key aspect of a randomized control trial that allows us to infer causality is randomization. If randomization is successful, then individuals selected for the treatment should be similar (on average) to those selected for the control. Therefore, a key component of any analysis of experimental data is to show that the treatment and control are similar based on observable characteristics.

To understand how a balance test might fail, let's imagine researchers are testing whether a given drug reduces cholesterol. The experimenters receive a list of 200 participants for the trial and decide the first 100 names on the list will receive the treatment and the second 100 will receive the placebo. What they fail to realize is that the list of applicants is sorted from youngest to oldest. Therefore, only the youngest participants receive the treatment. But younger applicants, on average, probably have lower cholesterol to begin with, so differences between treatment and control may be due to age, not due to the treatment.

If we perform a balance test (i.e. calculate average age by treatment and control) we would immediately detect our error. Now, no "real" experiment would ever be run in such a way. To do so would be ignoring standard protocol for well-run experiments! Still, it is important to understand whether there are differences between treated and control units before proceeding with the main analysis.Finding persistent differences with respect to observables could be evidence that randomization somehow failed. In our empirical application, we have information on the age of the student, so let's see if age varies between the treatment and control.

We want to retrieve the average age by Treatment Arm. To do so we can use a pivot table! The **Rows** will be the Treatment Arm, while the **Values** will be age. We can follow the exact same steps as in Section @ref(heading.pivottable)

Chapter 2

Intergenerational Mobility and Higher Education (Stata)

2.1 Intergenerational Mobility

In this section, we will be discussing the relationship between **intergenerational mobility** and the higher education system in the United Stata. Before we start exploring this relationship, we first need to understand the concept of intergenerational mobility. This concept asks a simple question: if you are born from low-income parents, what is the chance you move up the income distribution?

Generally, we think of societies with high rates of intergenerational mobility as more equal. If a society has high rates of intergenerational mobility, then that means individuals from low-income backgrounds have opportunities to move up the income distribution. One common path to upward mobility is the higher education system.

However, there are a few reasons why the higher education system may or may not lead to higher rates of intergenerational mobility. First, some colleges may not be particularly effective in increasing incomes. For example, for-profit colleges are often criticized as costly and low quality (See NYT Article). If a college accepts a large fraction of students from low-income backgrounds, it may still not promote intergenerational mobility if its students don't actually benefit from attending. Second, individuals from lower-income backgrounds may have lower access to the higher education system. Rising cost of college may be making this worse over time.

So how do we go about studying this question? We are going to use a dataset made publicly available by the researchers at **Opportunity Insights**. This data has made a big splash, both in academia and in policy circles.Figure 2.1 shows a headline from an Upshot article about the dataset. In this article, the authors discuss how this dataset has led us to new insights about how access varies dramatically at some schools. The headline reads, "Some Colleges Have More Students From the Top 1 Percent Than the Bottom 60. Find Yours". If a school has more students from the top 1 percent than the bottom 60, then it probably does not promote intergenerational mobility. Most of the students are already from relatively well-off backgrounds. This is our first hint that colleges may vary dramatically in the extent that they promote intergenerational mobility.

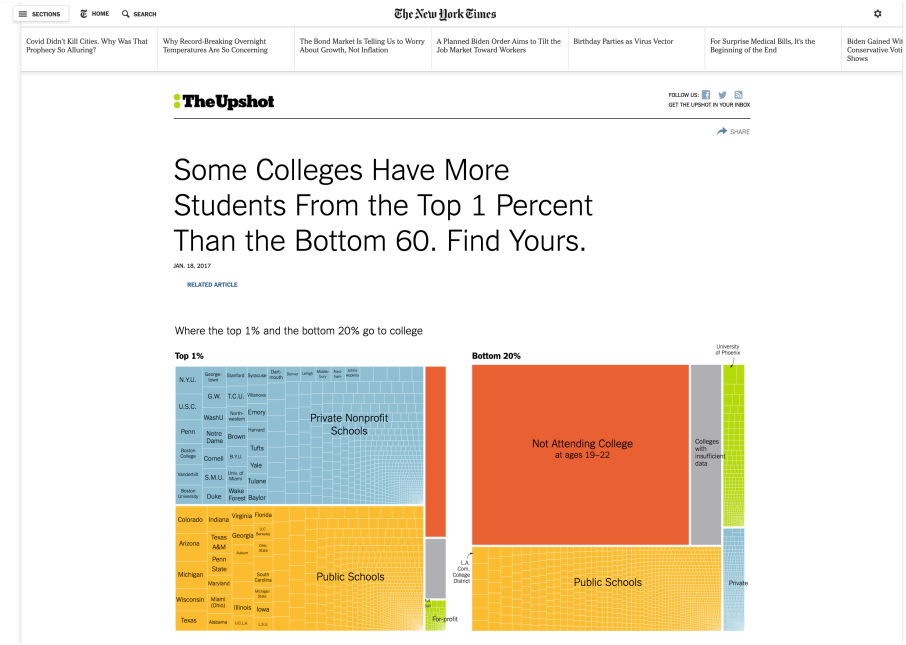


Figure 2.1: New York Times Graphic on Opportunity Insights

Let's get a little bit more into the details on what data we will be using in this empirical application. The data comes from federal income tax returns. Researchers at Opportunity Insights have linked to federal tax returns to tax records of universities and data from Pell grants. The result is a dataset with (1) Children linked to their parents (children are reported as dependents) (2) Income data for both children and parents and (3) children linked to the universities they attended.

This will give us everything we need to know to understand how a given college promotes intergenerational mobility. It will allow us to know both the know both

the background of students attending a given college as well as their eventual earnings. It is important to emphasize that this was a huge data task that gives us a new way to study an important question. This comprehensive data has only recently become available.

We have discussed at a high level the concept of intergenerational mobility. But when we turn to the data, we will need to compute a metric of intergenerational mobility that allows us to compare across colleges. Before introducing this measure, we will introduce a statistical concept that will be important throughout the course, and in particular, in understanding our measure of intergenerational mobility: **quantiles**.