Dora at MacIntosh Run, spring 2021

Temple of Heaven, Beijing – Qi’s son and parents

Trees above Ritika’s home

Day 2

Recap of Day 1

- **Data analytics** is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases
- Data mining is a **PROCESS** not a tool or formula.
- It starts with identifying what the **business issue** is and how data may be able to inform the decision.

Objectives for Day 2

- Problem finding and translating a **business issue** into a **data problem**
- Where do we get data and what does it look like?
- What is the “quality” of the data?
- How do we prepare the data for analysis using Excel?

- Good data analytics is good storytelling.

- There are many many decisions that are made about the data, even before the analysis begins.

- Few people even think about these decisions and don’t respect this preparatory work.

- But you have to be ready for those that are thinking and asking these questions.
- Being able to respond to these questions is critical to ensuring your audience believes what you discover.

- Often, you can tell your story is a single picture.

An example:

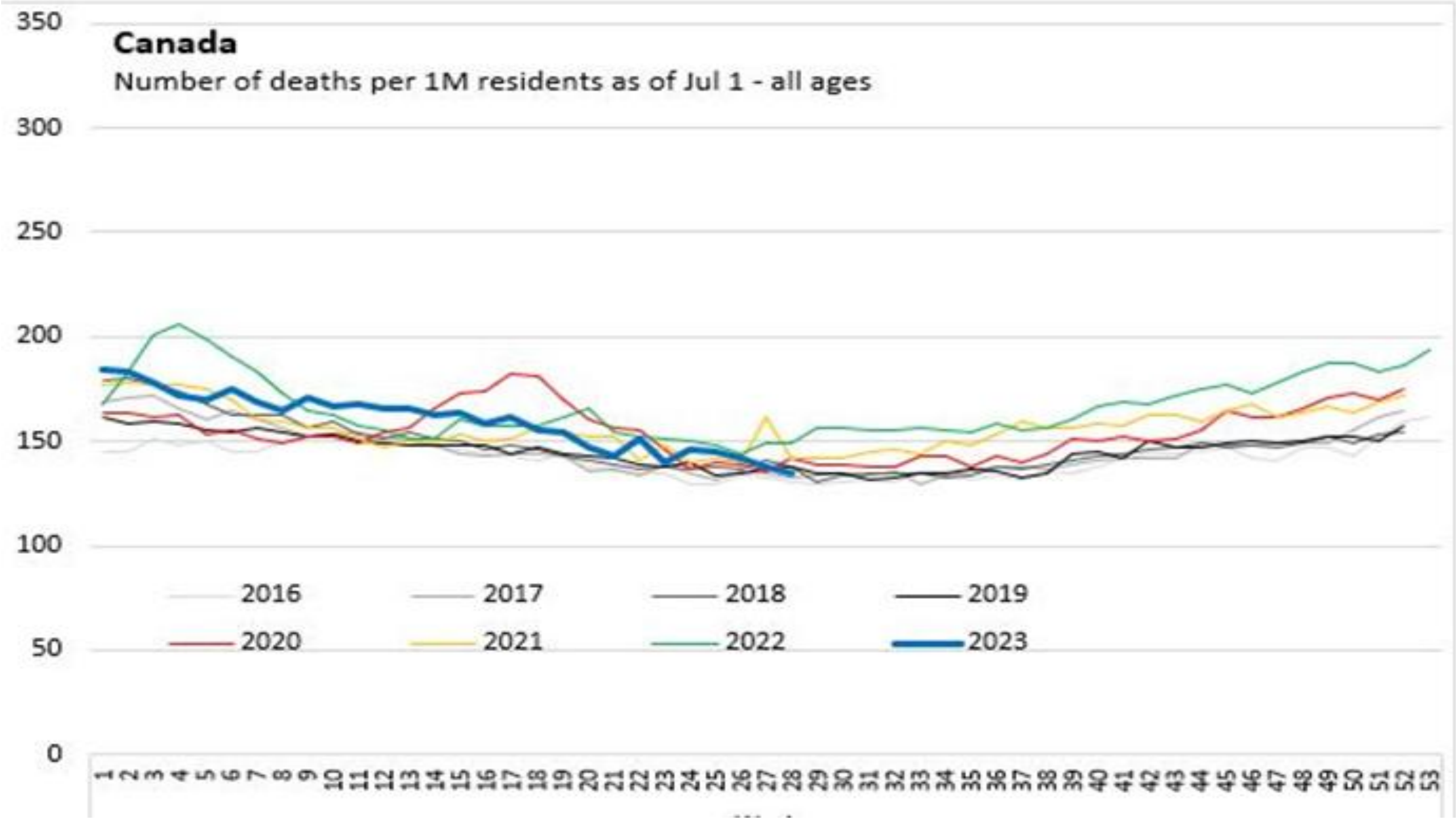
COVID has had a devastating impact on mortality in Canada

- but is this really true?

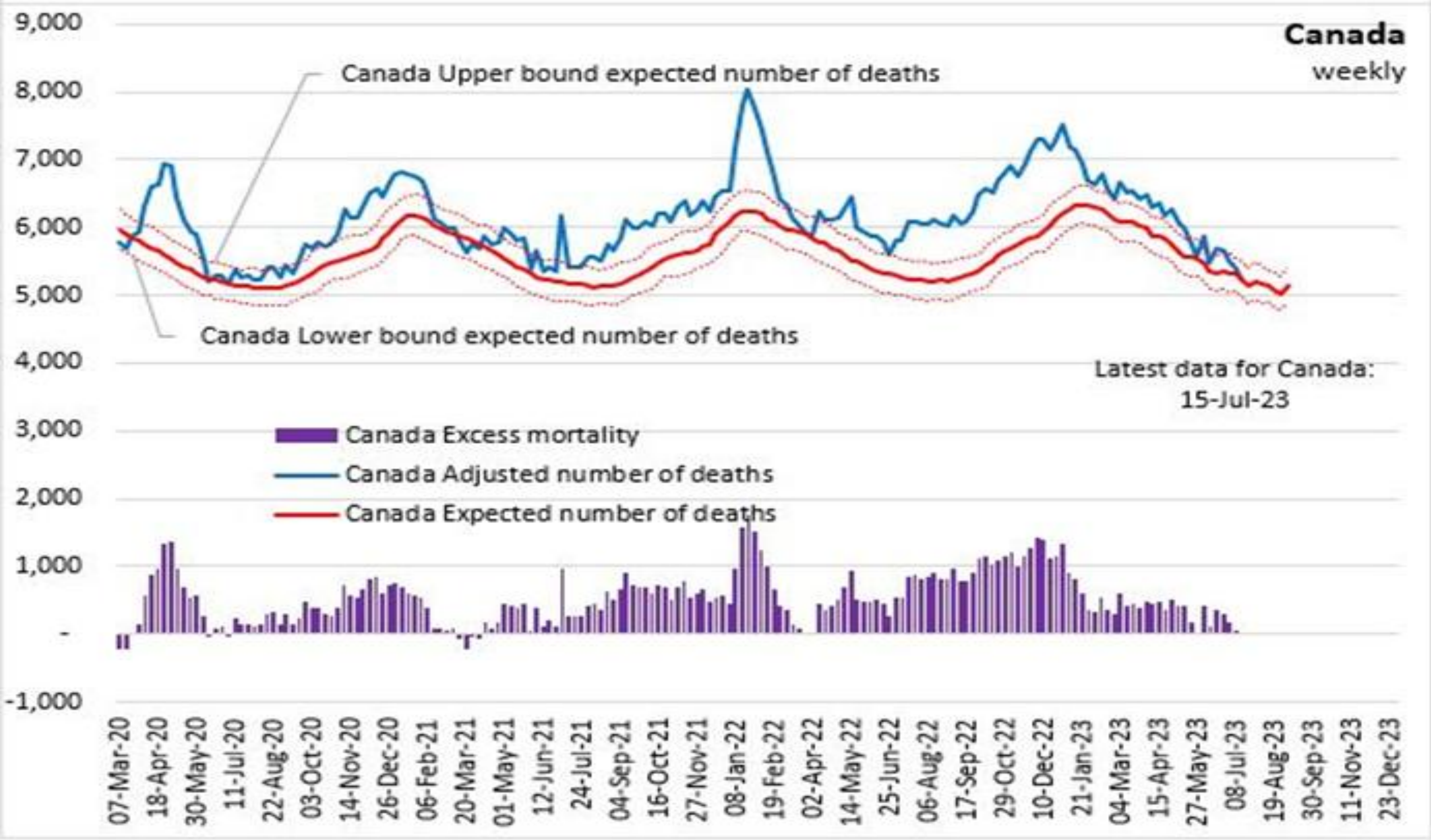
- How do you know a death was due to COVID?
- Most are old people who are going to die anyway.
- There is variation in mortality year to year. This is one of the up years and next year it will go down again.
- This is a hoax – fake news!

- What has mortality data looked like during the pandemic?

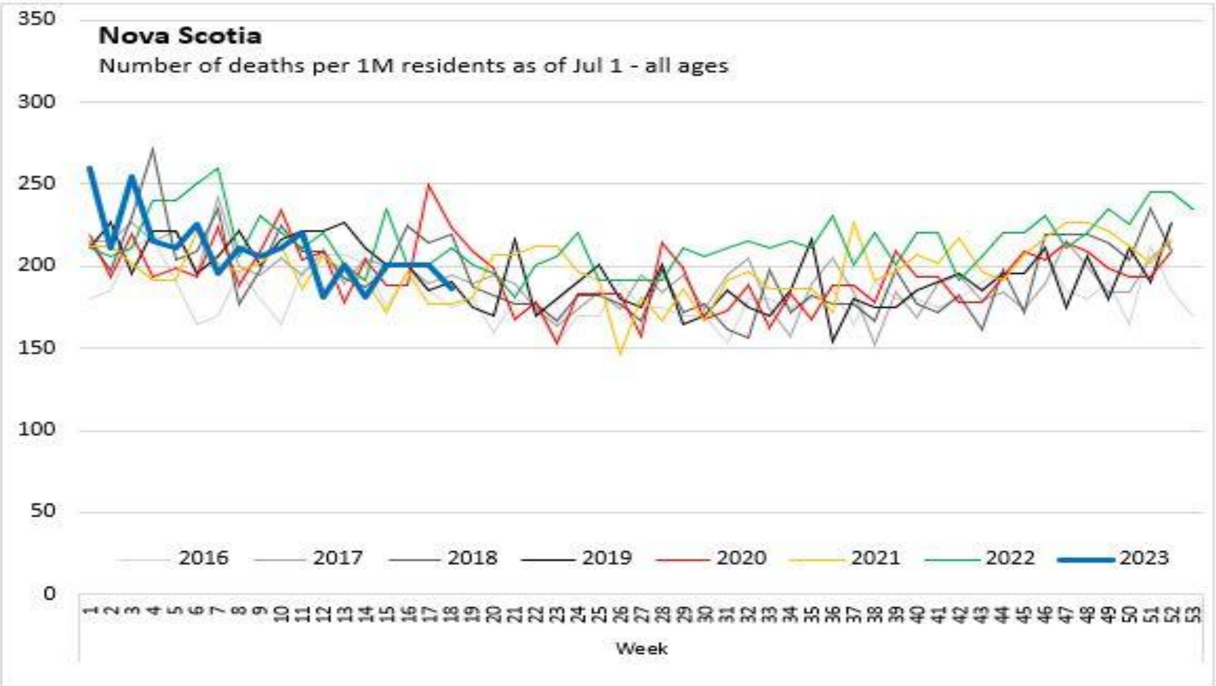
Source: Statistics Canada. Table 13-10-0768-01   Weekly death counts, by age group and sex; Table 13-10-0784-01 Adjusted number of deaths, expected number of deaths and estimates of excess mortality, by week

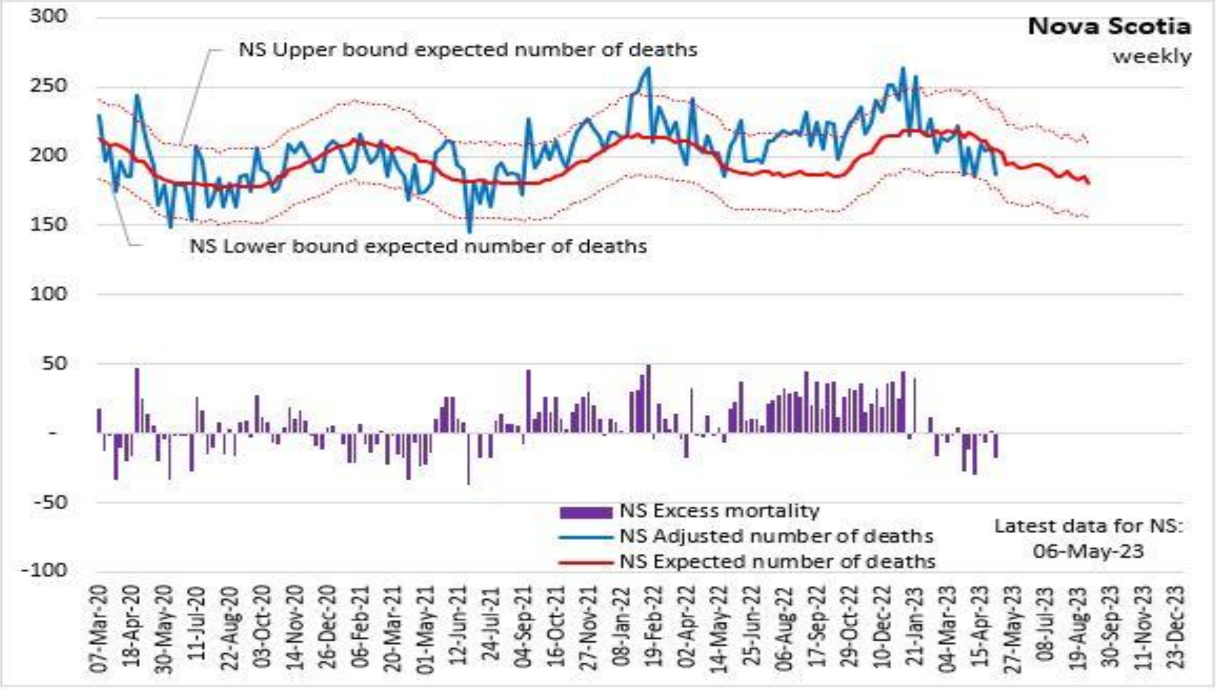


OK, but what is this supposed to show me? Mortality rates drop in summer?



Busy. Collapsed pre-COVID into a hi-lo range and put COVID timeline on top.





With a smaller population, weekly mortality rates are more volatile.

We did very well at the start of the pandemic, but since the arrival of Omicron and the lifting of restrictions, mortality rates have increased substantially.

But there is no longer the public support to re-introduce mandates – the model may say one thing, but it is important to understand all the actors for effective implementation.

But 2023 looks good – fewer deaths than expected! But maybe those that would have died in 2023, died in 2022!!

A simple picture can tell a powerful story.

Much thought has to go into

- what you are trying to say,
- how to measure it,
- how to “clean” the data,
- how to say it as simply as possible.
- Previous charts are from the DailyStats blog of the NS Department of Finance. They did not update the charts in 2024.

- The hard work is often boring.

- Now for the boring stuff …..

Business Understanding

- Often the organization has difficulty articulating their issue/problem.

- **Example:**
  - You are asked to prepare a report for a senior university administrator.
  - Students feel that the administration needs to understand how finances are affecting students.
  - But what does this mean? What is the issue?

A question might be:

“To what extent is the increasing cost of university education, high rents and limited student income affecting student academic and personal success and health?”

The question puts some boundaries on what is in and not in the report.

It may also be seen as “biased” because attention is drawn to costs *increasing* and student income being *limited*.

- The report should be structured around answering very specific questions.

- What would be good questions that would speak clearly to the problem you have defined?

- Report might focus on answering specific questions:
    - What are sources of student income/financial resources?
    - How does income vary among student populations?
    - Who is working while at school and how much?
    - Does working affect mental health?
    - Who is stressed about paying for school?
    - How is working and/or stress affecting grades?

- Many questions may need to be asked before you can define a problem that we can analyze.

Data Understanding

- Where can we get data?
  - Administrative databases (e.g., Banner has student records)
  - 3rd party sources (Statistics Canada, purchased data,…)
  - a course survey?

  - Regrettably, few groups are looking seriously at this issue.

Samples and Populations

- A **population** is “everybody” whereas the **sample** is who you have data about.
- You may wish to talk about consumers, but you only have data about those who have purchased your service/product.
- Ideally we want to talk about a general population, but this can only be done if our sample is representative of the population.

- The only way to ensure your sample is representative is if it is a **“random” sample**.
- To be a random sample, everyone in the population must have equal chance of being included in your sample. An example of this is seen later when we introduce **“experiments”.**

Administrative data

- Structured data databases – customer data, transaction data, supplier data, …
- Databases are complex interconnected files (**relational database**)
- Amazon has records about me.
  - In one file they have the various addresses that I have shipped to.
  - In another they have the various credit cards I have used.
  - In another there are the many purchase transactions.
  - In another there is the history of what items I have viewed and when, how long,…

Extract, Transform, Load (ETL)

- To do analysis, we need to **extract** the data we want and organize it in a fashion we can easily analyze.
- Usually this means copying data into a “**flat file**” that looks like a simple spreadsheet.
- Each row is one “**observation**” (**record**) and the columns are the **attributes (variables)**.
- If we were studying customers, then ideally we would want one row (record) for each unique customer.

- Since I have made many visits and made many purchases, I might have to **transform** the raw information into new variables, such as,
  - Average number of visits per month;
  - Average purchase in $$;
  - Average number of items purchased per order;
  - Where I live (billing address?).

- When we have the data is the desired form, we will **load** it for analysis.

- We want a **FLAT** file. Each column represents a variable. Each row is an observation.

Open the **CoffeeShop_LoyaltyData_LineItems** file on TopHat

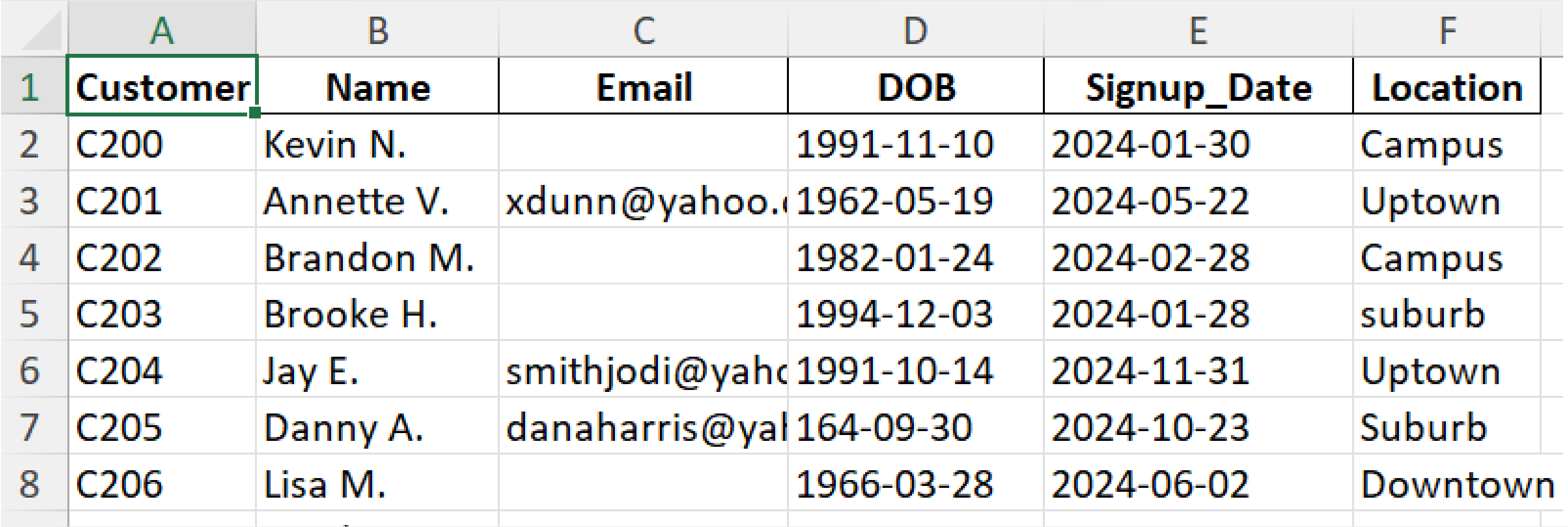
There are two sheets that represent 2 files that would be in Perk’s customer database.

One has customer loyalty card information.

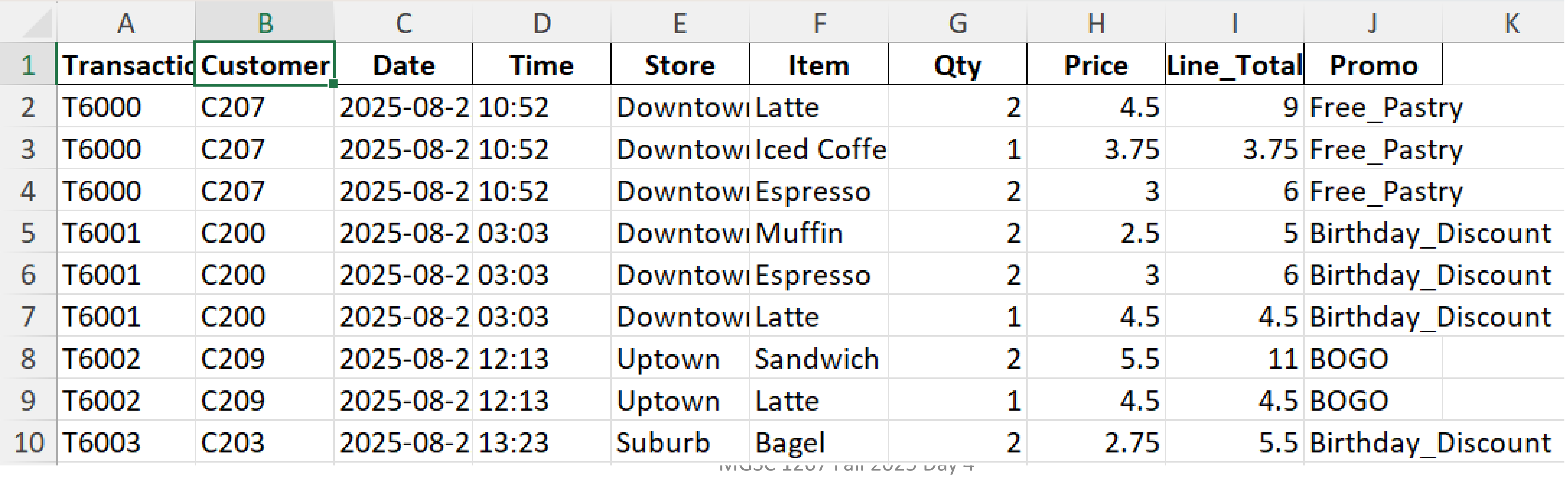
The second has individual transaction data.

What issues do you see in preparing a data file to explore opportunities to improve sales?

Perk’s Café



	A	B	C	D	E	F
1	Customer	Name	Email	DOB	Signup_Date	Location
2	C200	Kevin N.		1991-11-10	2024-01-30	Campus
3	C201	Annette V.	xdunn@yahoo.(	1962-05-19	2024-05-22	Uptown
4	C202	Brandon M.		1982-01-24	2024-02-28	Campus
5	C203	Brooke H.		1994-12-03	2024-01-28	suburb
6	C204	Jay E.	smithjodi@yahc	1991-10-14	2024-11-31	Uptown
7	C205	Danny A.	danaharris@yal	164-09-30	2024-10-23	Suburb
8	C206	Lisa M.		1966-03-28	2024-06-02	Downtown
9						



	A	B	C	D	E	F	G	H	I	J	K
1	Transactic	Customer	Date	Time	Store	Item	Qty	Price	Line_Total	Promo	
2	T6000	C207	2025-08-2	10:52	Downtowı	Latte	2	4.5	9	Free_Pastry	
3	T6000	C207	2025-08-2	10:52	Downtowı	Iced Coffe	1	3.75	3.75	Free_Pastry	
4	T6000	C207	2025-08-2	10:52	Downtowı	Espresso	2	3	6	Free_Pastry	
5	T6001	C200	2025-08-2	03:03	Downtowı	Muffin	2	2.5	5	Birthday_Discount	
6	T6001	C200	2025-08-2	03:03	Downtowı	Espresso	2	3	6	Birthday_Discount	
7	T6001	C200	2025-08-2	03:03	Downtowı	Latte	1	4.5	4.5	Birthday_Discount	
8	T6002	C209	2025-08-2	12:13	Uptown	Sandwich	2	5.5	11	BOGO	
9	T6002	C209	2025-08-2	12:13	Uptown	Latte	1	4.5	4.5	BOGO	
10	T6003	C203	2025-08-2	13:23	Suburb	Bagel	2	2.75	5.5	Birthday_Discount	

- For the report on student finances, we could extract data from Banner

  - Total fees in current term
  - Current balance owing on student account
  - Does student have a student loan? Amount?
  - Does student have scholarship or bursary? Amount?
  - Has student applied for a bursary?
  - What is the student’s cumulative GPA? What was the GPA in the last term?
  - Is the student Canadian or on student visa?
  - If Canadian, is the student from Halifax or away?
  - If international, what country is the student from?

- Although we can get a lot of data, we don’t know anything about what the student does outside school or how they feel about their situation.

- Frequently, we can get some of the data we would like, but not always in the form we want it.

- You work with what you have and explore how you can address what is missing.

- The resulting data file may be a combination of data from multiple sources that have been merged. Easy? Not! Can present unintended problems.

Surveys

- To capture data not available in the administrative databases, you may try doing a survey.
- You invite individuals to answer a variety of questions
- In previous terms, students in MGSC 1207 were asked how much they earned and saved in the summer, how much they were paying in tuition, whether they were worried about paying for school, what was their GPA, had they ever failed a course, where were they from,….
- Some questions corresponded to what might have been in Banner, but many did not.

What if we wanted to merge the survey and administrative data sets

What issues might we encounter?

- Missing data
- Privacy/permission
- How do you link anonymous survey data with personal administrative data?
- What if the data is different?

Challenges with data

- Only 75% of students participated in the 2019-2024 surveys.
- Many online surveys are open to an unlimited number of people. Maybe 1,000 respond but that might be less than 0.1% of those who could have.

- Surveys are an example of an "**observational study**" – you only have data on those you have data for. What about those who did not respond? Would their responses have been similar? Don't know since they didn't tell you.

- **Administrative data is also observational**. In studying student retention and persistence, we only have data on those students who enrolled at Saint Mary's. What about those who applied but did not enroll? What about those that never applied?

Experiments vs Observational Studies

- How can we control for some challenges that come with an observational study?

- In online marketing, they often do **A/B experiments**.
  - When you visit a webpage, you may see a different advertisement than another visitor.
  - Visitors are randomly shown ad A or ad B.
  - Your response is tracked and the firm evaluates which ad generates a more positive response.
  - With experiments you can exclude all other factors that may affect behavior so you can be confident in drawing conclusions that can be generalized to the general population.

- Experiments control for factors that allow you to generalize.
- Unfortunately, most of the time you cannot randomly assign “treatments” and you must work with observational data.

Data Quality

- Data quality may be more important than the volume of data.

- What is data quality?

- In 2010, students were asked how much they earned last summer. How much did you earn? Do you remember? Most gave numbers to the nearest $1,000 but one said $9,655. How can he be so precise? One said $650,025!
    - Some cannot remember with precision
    - Some will make data entry errors
    - Some are not honest
    - Some will misunderstand the question

- The problem is not just with surveys. There can be errors in administrative data.
- Many fall students show as transfer students on Banner, but when I look at the transcript, there are no transfer credits. Was no credit given or are they not transfers?

2022 Canadian Student Wellbeing Survey

Background

Universities and colleges around the world partner with Studiosity to provide academic writing and core skills support 24/7, connecting students to help when they need it.

They currently work with 75% of universities in Australia, 21 universities in the United Kingdom, and have recently started providing support to Canadian institutions.

As of 2022, Studiosity's services are available worldwide to over 1.6 million students.

Research objectives

- This annual survey seeks to better understand the motivations, emotions, and demands of postsecondary students in Canada, and how to better tailor initiatives and solutions. The survey investigates key areas of student wellbeing.
- In 2022, the survey focused on topics such as experiences of stress, importance of grades, motivation and, cheating, feelings towards the future, intent to withdraw, engagement with institution, and the transition to postsecondary.
- In addition, the results were benchmarked with the 2021 survey to better understand how behaviours and attitudes have changed and how these themes have shifted after another year of the COVID-19 pandemic.

Methodology & Sample

- Studiosity produced the questions for this survey, and Angus Reid Forum gathered the responses from current Canadian postsecondary students via an online survey.
- The survey ran from March 10 to March 24, 2022 and gained a total of 1,014 responses.
- The sample frame was balanced to ensure representation of men and women in proportion to their overall share of the Canadian postsecondary student population (56% female, 43% male), as well as to ensure statistically significant representation from different regions of the country.
- The sample was comprised of 75% full-time students and 25% parttime students, and was conducted in English and French.

What are your thoughts? **How did they obtain their sample?**

Who is the population?

Do you think the sample frame aligns well with the population?

Who was in the sample frame? Who had the opportunity to participate?

What cautions might you have about accepting the survey results?

Data quality questions to ask

- Is the data really what you think it is (data definition)?
- Is it coded consistently?
- Are there coding errors? How can you tell?
- Is there missing data?
  - Is that important?
  - Should you delete the record?
  - Guess the missing value?
- Are there duplicates?
- Are the values “reasonable”?

Often our biggest issue is the data that we do not have.

- In a survey, some respond and many refuse – the **Silent Majority**.
- Are non-respondents different from respondents? Does this matter for our study?

- Regardless of whether we have administrative data, survey data or experimental data, there may be missing values for some variables.
- Causes are numerous:
    - In surveys, fatigue leads to respondents stopping part way through.
    - With transaction data, customers only buy a subset of products available.
    - In healthcare data, we only have lab results for the tests that were performed.

- Non-respondents are a challenge. If we know that 50% of the population is female and 68% or respondents are female, we can apply weights to female responses to lower the “value” of each female response and increase the value of each male response.
- But there are limits to what can be done with weighting.

- For missing values, what do you do?
  - Drop records with missing data?
  - Guess at the value of the missing item? How?
    - Use the average? Or median?
    - Build a model to predict the missing value? Regression, KNN, …
  - Drop the variable if too many values missing?

Summary

- Business problems are often poorly defined and unstructured.
- To create a data analysis problem from a business problem you must define questions that you can measure.
- Often you have data for some questions, but maybe not all, or the data may not be exactly what the question calls for.
- Your data may come from an administrative database, a survey, a 3rd party, or a combination of all of these if you can link the records together.
- Most of the time you are using “observational data” – what you observed in an uncontrolled environment. This may limit your ability to generalize to a broader population.

Summary cont’d

- Ideally we would like to have controlled experiments, but often people will not let you do experiments on them.
- There are usually some data quality challenges, even with admin data.
- Many valuable insights can only be obtained through surveys – surveys present many data quality challenges – non-response, bias, data errors, honesty, misunderstanding,…
- Data quality problems may be addressed through various methods, but do they bias the sample in new ways?

Dora at Herring Cove Provincial Park - spring 2021

Mera loves scuba diving

Mina loved the northern lights
when she lived in Saskatchewan

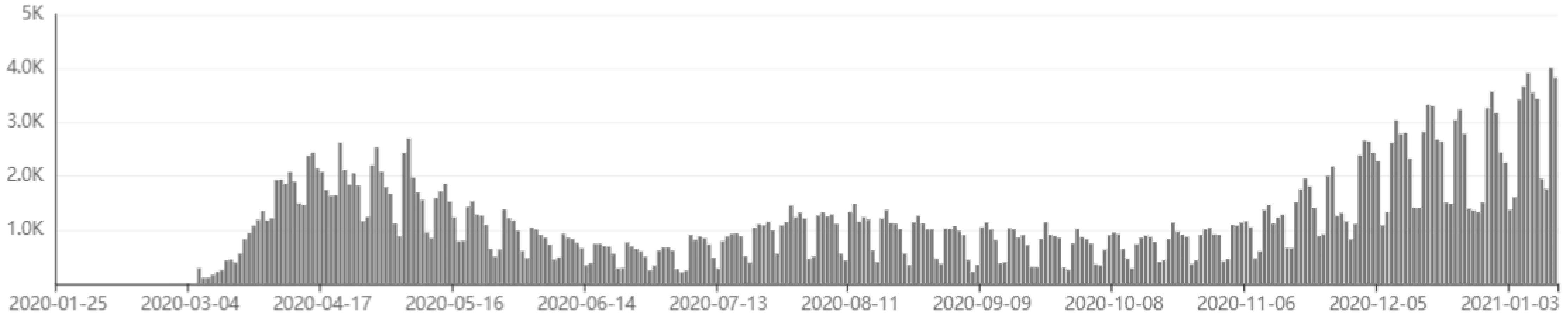
Objectives for Day 2 part 2

- Some basic Data Preparation using Excel
- Protecting your data
- What do the variables mean? Data dictionary
- Too much data - Selecting variables
- Renaming variables for easier recollection
- Recoding values for easier interpretation (VLOOKUP exact match)
- Recoding values into groups (VLOOKUP approximate match)

What does data look like?

- Consider (raw) data that you may see often in the media.
- COVID case and fatality data is readily available on many sites.
- Google and MSN have attractive sites with good graphics.
- Coronavirus (COVID-19) (msn.com)
- https://news.google.com/covid19/map?hl=en-CA&gl=CA&ceid=CA%3Aen

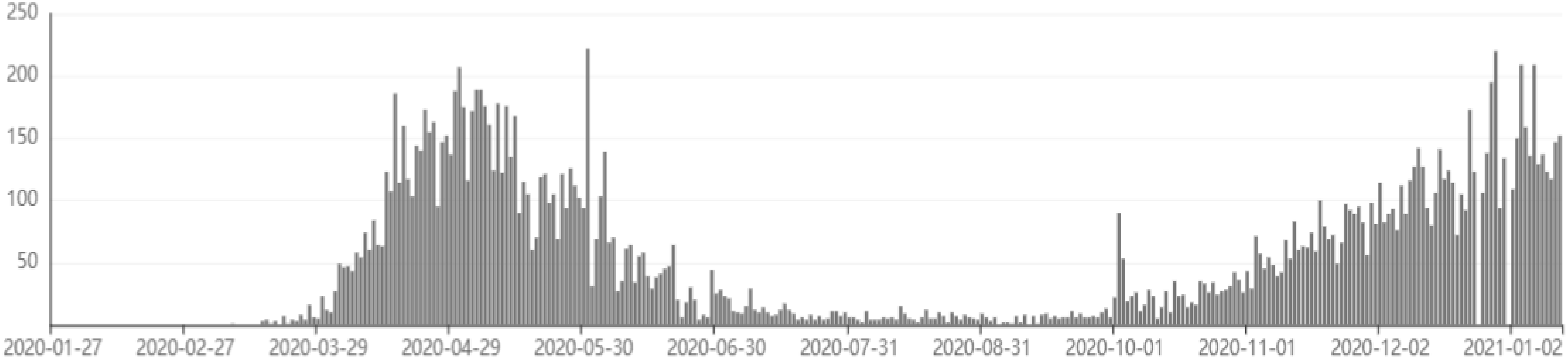
USA daily fatalities from MSN



- This was just as they entered their 3rd wave.

- But there seems to be a strange “toothy” look.
  Why?

Daily fatalities in Canada (MSN)



- This is the comparable data for Canada.
- Better but strange?

- We are more often **consumers** than producers of data or summaries from data.
- Think of where the data comes from.
- Be critical consumers.

- Let us look at the producer’s situation.

Business Issue

What are student expectations of earnings 2 years after they graduate?

- Are expectations similar among students?
- Do expectations vary by program?
- Are expectations different for domestic versus international students?
- Are expectations similar between males and females?
- Do student expectations change as they progress through their program?
- Are students concerned about finding a job?
- Are salary expectations different if a student is concerned compared to those not concerned?

Data?

- winter of 2010, a national survey of undergraduate students
- We have the data for 811 Saint Mary's students who participated.

q14a Indicate how concerned you are about finding any job after graduation.

q14b Indicate how concerned you are about finding a job in your field of study.

q14c Indicate how concerned you are about finding a job that will pay a salary you desire.

q14d Indicate how concerned you are about finding a job in a place you want to live.

q15 What do you expect your annual salary to be two years after finishing your post-secondary education?

q16 How long do you think it will take you to find a job after completing your degree?

Data cont’d

- The previous data can described as “outcomes”.
- The questions asked how these outcomes might differ among various groups of students.
- We have the following demographic data.

q3 In what year did you first enrol in your current institution?
q5 In what type of program are you currently enrolled?
q7 In which province or territory did you complete high school?
q53 Are you: Male/Female
q54 How old are you?
q56 What was the highest level of education completed by either of your parents?
q57 What language do you generally speak when at home with your family?

Excel

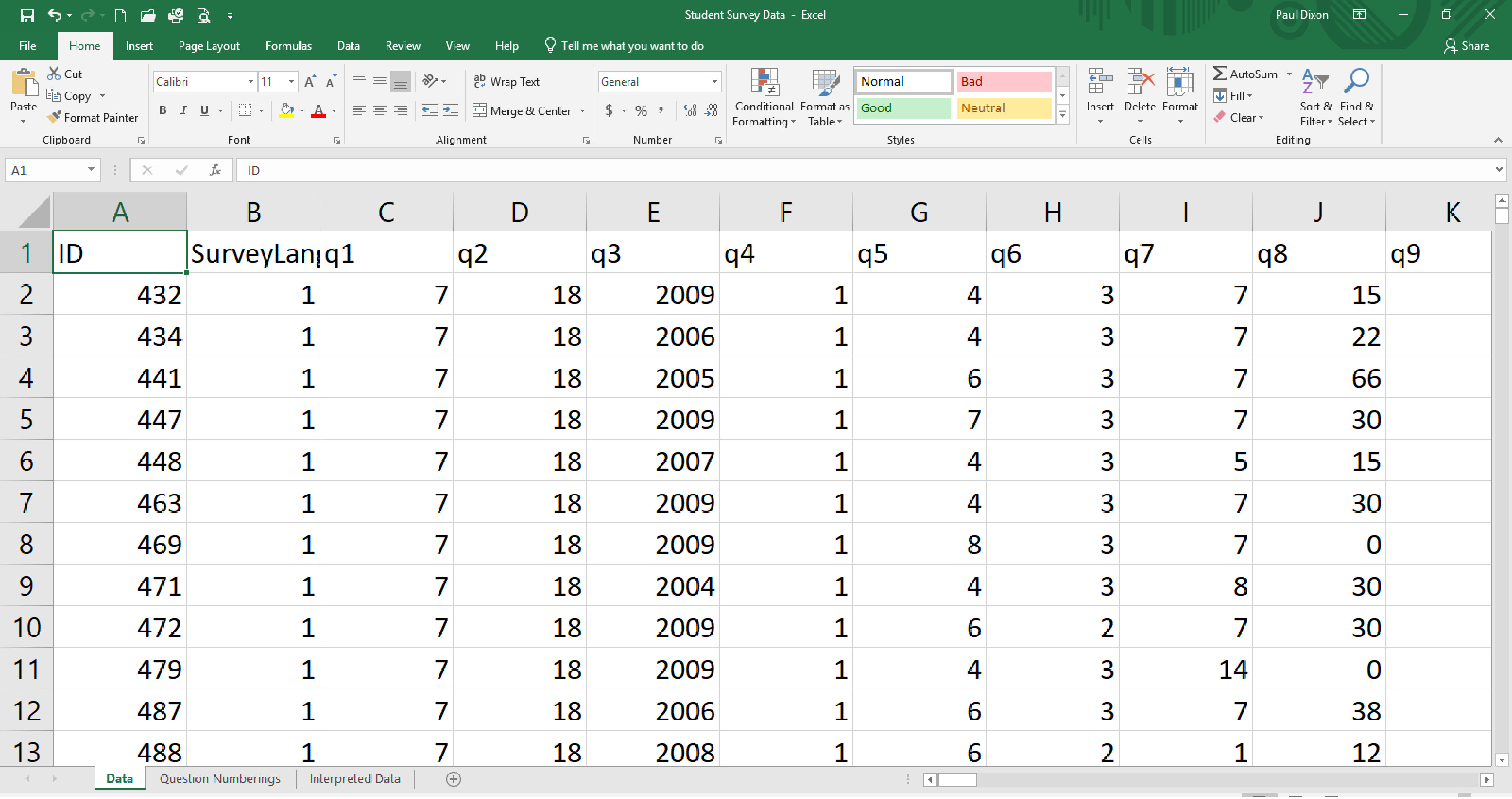
- Open the file “Chapter 2 Student Survey Data 2010”
- You will see a “workbook” with 3 sheets (tabs) at the bottom.
- There are other tabs that reflect actions we will take today.

- Click on the Data sheet.

- You should see a grid with many columns labeled with letters A, B, C, D, …. and rows that are numbered 1, 2, 3,…

- The first row has headings: ID, SurveyLanguage, q1, q2, q3, ….

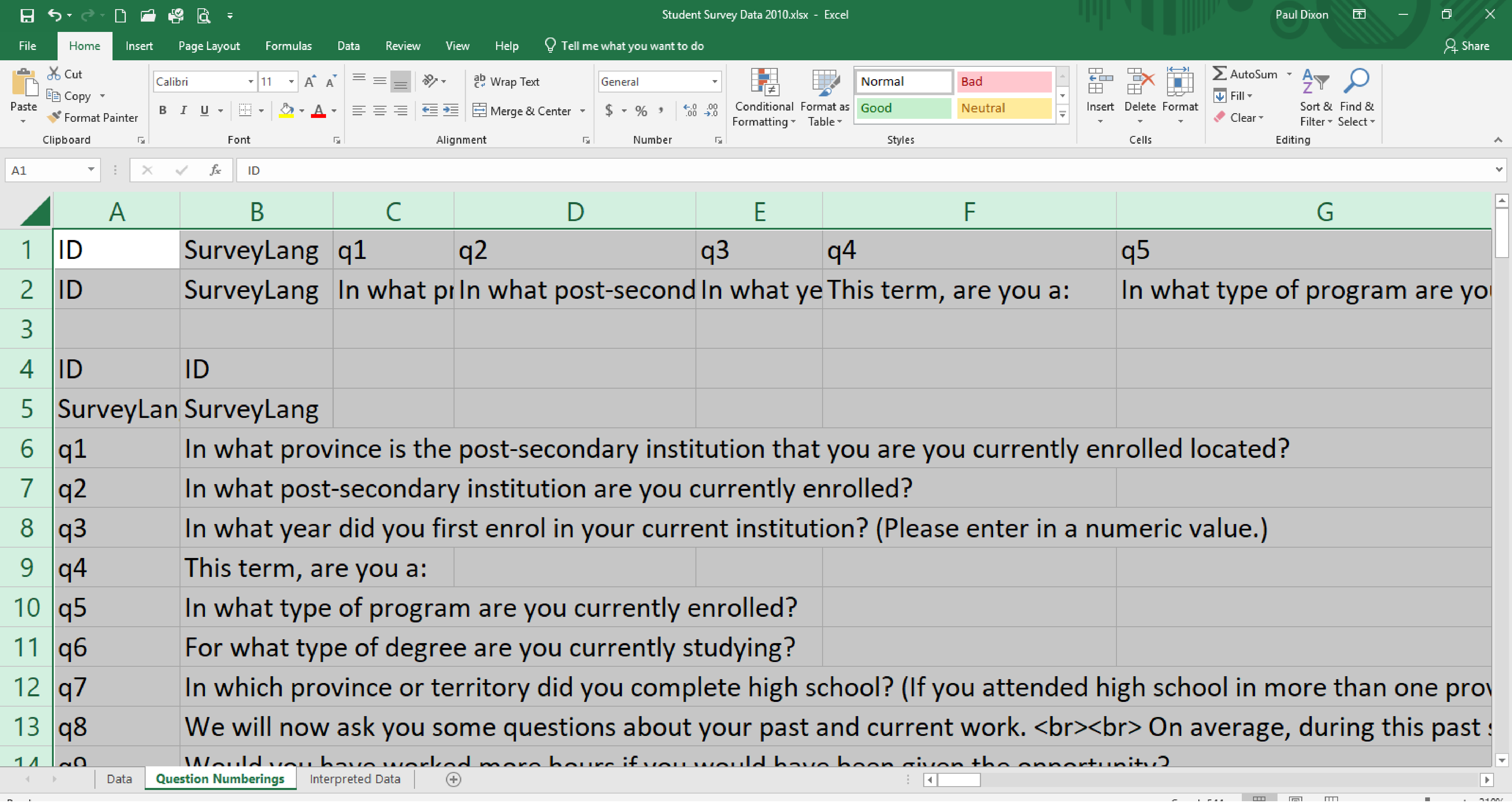
- The headings describe what is in each column. Each row represents a student’s response to the survey.



	A	B	C	D	E	F	G	H	I	J	K
1	ID	SurveyLan	q1	q2	q3	q4	q5	q6	q7	q8	q9
2	432	1	7	18	2009	1	4	3	7	15	
3	434	1	7	18	2006	1	4	3	7	22	
4	441	1	7	18	2005	1	6	3	7	66	
5	447	1	7	18	2009	1	7	3	7	30	
6	448	1	7	18	2007	1	4	3	5	15	
7	463	1	7	18	2009	1	4	3	7	30	
8	469	1	7	18	2009	1	8	3	7	0	
9	471	1	7	18	2004	1	4	3	8	30	
10	472	1	7	18	2009	1	6	2	7	30	
11	479	1	7	18	2009	1	4	3	14	0	
12	487	1	7	18	2006	1	6	3	7	38	
13	488	1	7	18	2008	1	6	2	1	12	

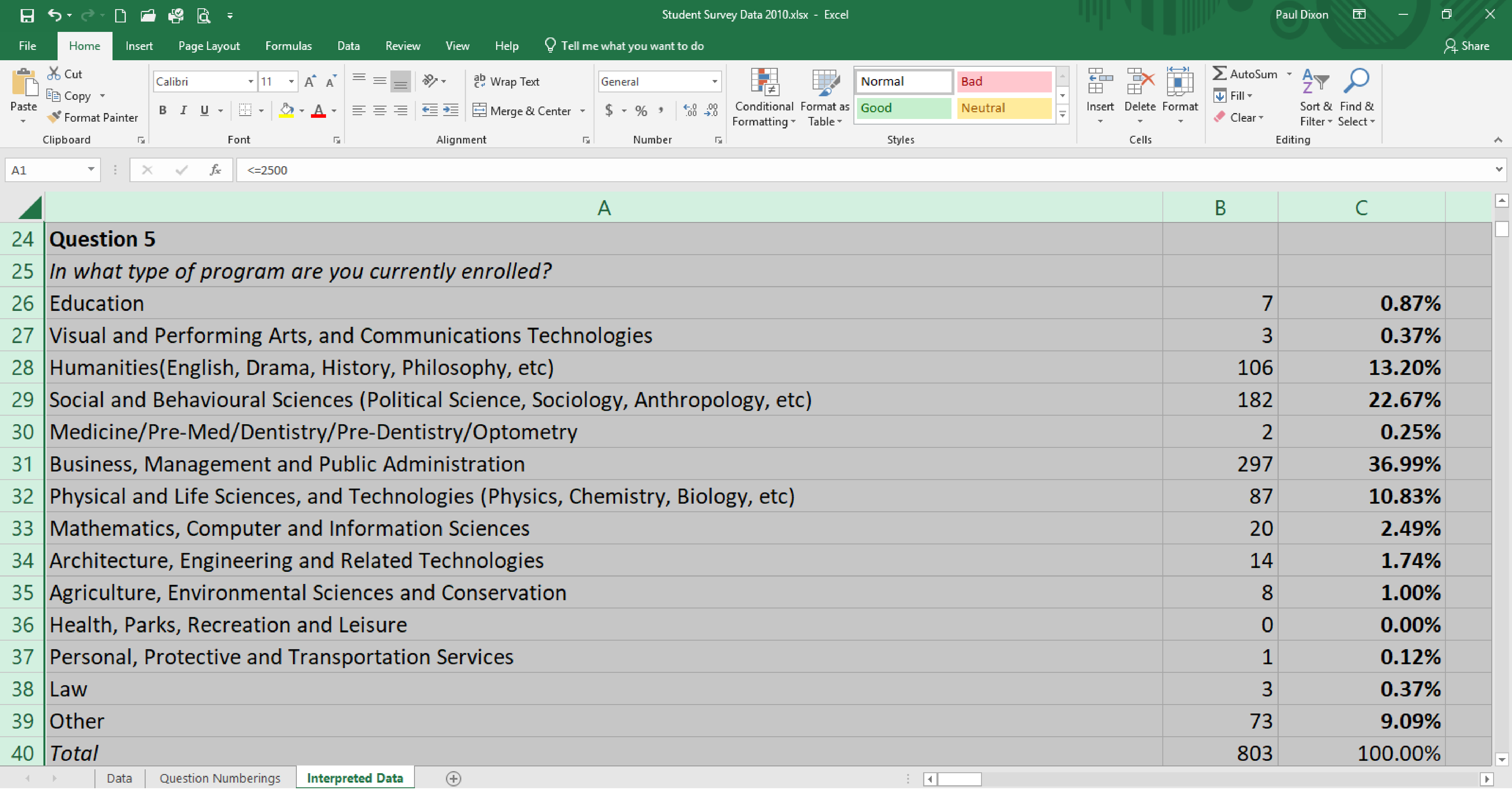
- How do we know what q1 is?
- Previously, I stated that q5 was the type of program. But we see numbers. What program is 4?

- Every data file needs a **Data Dictionary**.
- A data dictionary explains what each variable represents and how to interpret the data.
- This file has a tab labeled **Question Numberings**.



	A	B	C	D	E	F	G
1	ID	SurveyLang	q1	q2	q3	q4	q5
2	ID	SurveyLang	In what pr	In what post-second	In what ye	This term, are you a:	In what type of program are yo
3							
4	ID	ID					
5	SurveyLan	SurveyLang					
6	q1	In what province is the post-secondary institution that you are you currently enrolled located?					
7	q2	In what post-secondary institution are you currently enrolled?					
8	q3	In what year did you first enrol in your current institution? (Please enter in a numeric value.)					
9	q4	This term, are you a:					
10	q5	In what type of program are you currently enrolled?					
11	q6	For what type of degree are you currently studying?					
12	q7	In which province or territory did you complete high school? (If you attended high school in more than one prov					
13	q8	We will now ask you some questions about your past and current work. On average, during this past s					
14	q9	Would you have worked more hours if you would have been given the opportunity?					

- Somewhat useful, but it doesn’t tell me what the answers should look like.
- For q5, it still doesn’t tell me what 4 represents
- A good data dictionary should tell me what the variable represents and how to understand the values.
- Look at the tab **Interpreted Values**



A1 | <=2500

	A	B	C
24	**Question 5**		
25	*In what type of program are you currently enrolled?*		
26	Education	7	**0.87%**
27	Visual and Performing Arts, and Communications Technologies	3	**0.37%**
28	Humanities(English, Drama, History, Philosophy, etc)	106	**13.20%**
29	Social and Behavioural Sciences (Political Science, Sociology, Anthropology, etc)	182	**22.67%**
30	Medicine/Pre-Med/Dentistry/Pre-Dentistry/Optometry	2	**0.25%**
31	Business, Management and Public Administration	297	**36.99%**
32	Physical and Life Sciences, and Technologies (Physics, Chemistry, Biology, etc)	87	**10.83%**
33	Mathematics, Computer and Information Sciences	20	**2.49%**
34	Architecture, Engineering and Related Technologies	14	**1.74%**
35	Agriculture, Environmental Sciences and Conservation	8	**1.00%**
36	Health, Parks, Recreation and Leisure	0	**0.00%**
37	Personal, Protective and Transportation Services	1	**0.12%**
38	Law	3	**0.37%**
39	Other	73	**9.09%**
40	*Total*	803	100.00%

- Not ideal, since it still doesn’t say what 4 is.
- But the 4th entry is Social and Behavioral Sciences.

- Look at some of the other entries.
- Do we teach Visual and Performing Arts? Education? Medicine?
- Hmmm. We will look at the actual data later.

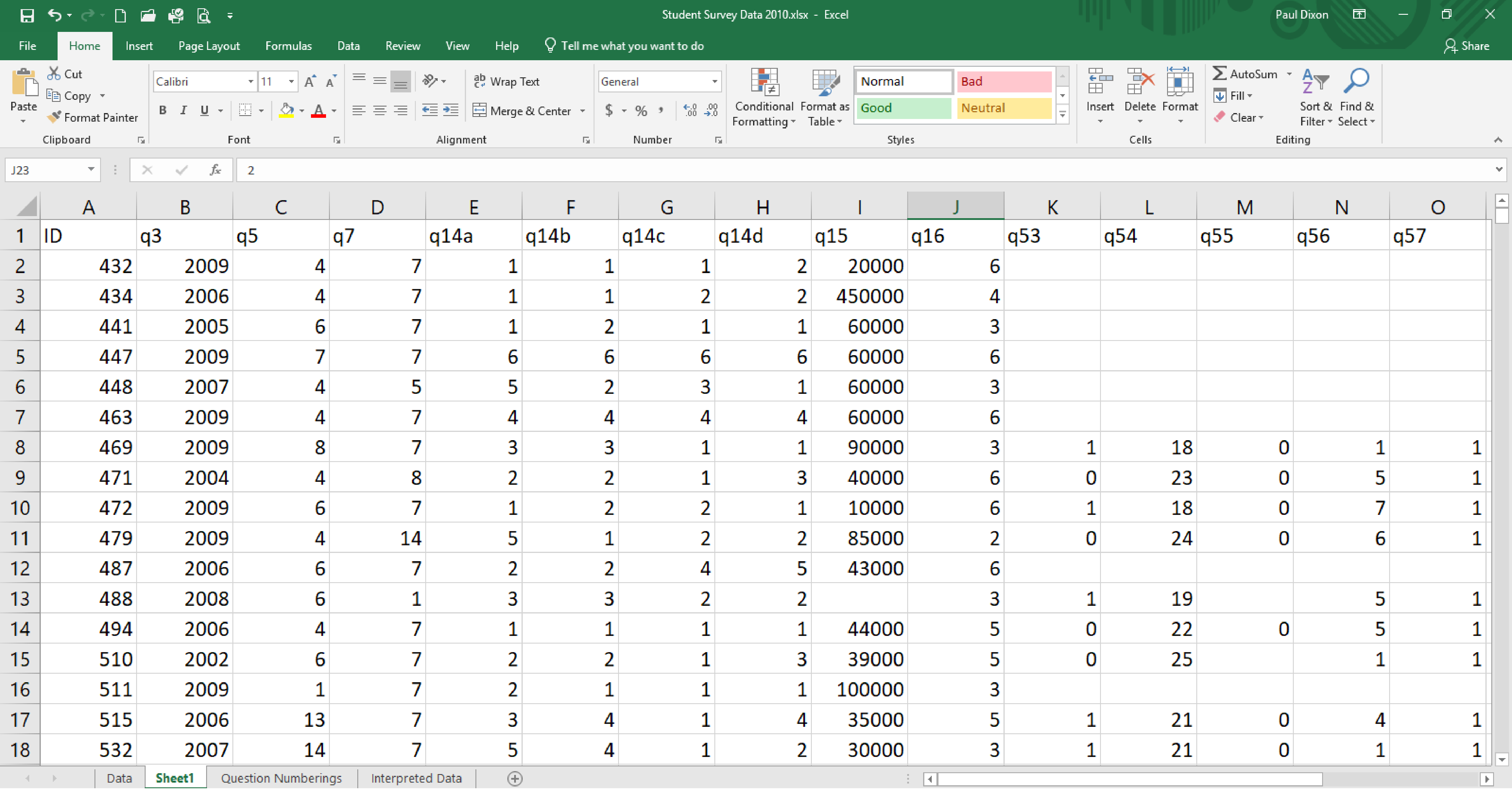
- There are a lot of variables. Do we need all of them?
- The headings are hard to understand.
- The values are hard to understand. Need to look up what the numbers mean.

- We will likely want to modify the data file.
- Protect your data.
  - Once you start changing it, you may not be able to go back.
  - What if you make a mistake?

- Most data analysis software maintains a script of changes (like a program) that is separate from the data. You never actually change the data file.
- **Excel makes changes as you make them. Dangerous.**
- **Make a copy of your data and only edit the copy.**

- Copy the questions of interest to a new sheet.
- Select the desired columns (one at a time or in blocks) then Copy and Paste.

ID	Record number
q3	In what year did you first enrol in your current institution?
q5	In what type of program are you currently enrolled?
q7	In which province or territory did you complete high school?
q14a	Indicate how concerned you are about finding any job after graduation.
q14b	Indicate how concerned you are about finding a job in your field of study.
q14c	Indicate how concerned you are about finding a job that will pay a salary you desire.
q14d	Indicate how concerned you are about finding a job in a place you want to live.
q15	What do you expect your annual salary to be two years after finishing your post-secondary education?
q16	How long do you think it will take you to find a job after completing your degree?
q53	Are you: Male/Female
q54	How old are you?
q56	What was the highest level of education completed by either of your parents?
q57	What language do you generally speak when at home with your family?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	q3	q5	q7	q14a	q14b	q14c	q14d	q15	q16	q53	q54	q55	q56	q57
2	432	2009	4	7	1	1	1	2	20000	6					
3	434	2006	4	7	1	1	2	2	450000	4					
4	441	2005	6	7	1	2	1	1	60000	3					
5	447	2009	7	7	6	6	6	6	60000	6					
6	448	2007	4	5	5	2	3	1	60000	3					
7	463	2009	4	7	4	4	4	4	60000	6					
8	469	2009	8	7	3	3	1	1	90000	3	1	18	0	1	1
9	471	2004	4	8	2	2	1	3	40000	6	0	23	0	5	1
10	472	2009	6	7	1	2	2	1	10000	6	1	18	0	7	1
11	479	2009	4	14	5	1	2	2	85000	2	0	24	0	6	1
12	487	2006	6	7	2	2	4	5	43000	6					
13	488	2008	6	1	3	3	2	2		3	1	19		5	1
14	494	2006	4	7	1	1	1	1	44000	5	0	22	0	5	1
15	510	2002	6	7	2	2	1	3	39000	5	0	25		1	1
16	511	2009	1	7	2	1	1	1	100000	3					
17	515	2006	13	7	3	4	1	4	35000	5	1	21	0	4	1
18	532	2007	14	7	5	4	1	2	30000	3	1	21	0	1	1

- It is common practice to have some type of record indicator so that you can uniquely identify each response.

- Give the sheet a name, such as **Grad Expectations**.

- Excel has special capabilities for data files.
- Go to tabs at the top of the screen, and select **Insert**.
- Select **Table**

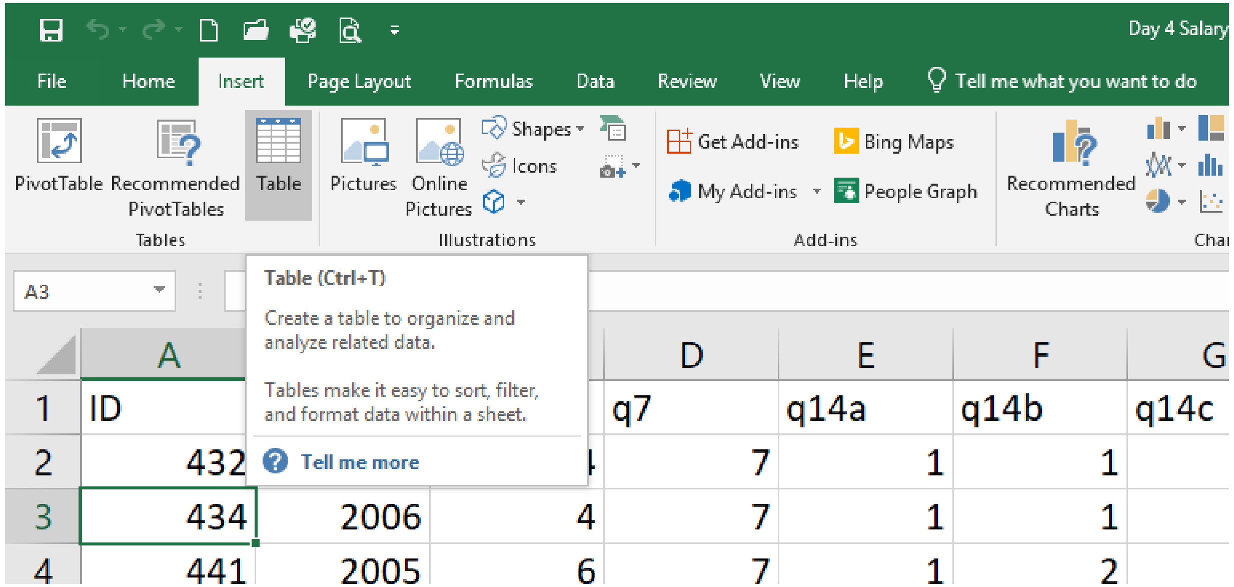


Table (Ctrl+T)

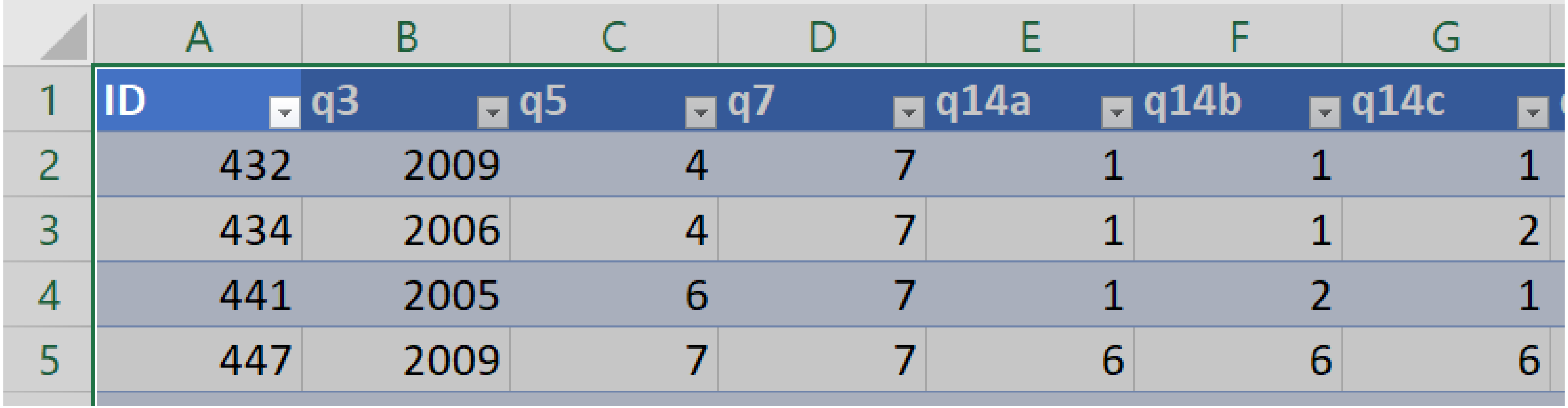
Create a table to organize and analyze related data.

Tables make it easy to sort, filter, and format data within a sheet.

Tell me more

	A	B	C	D	E	F	G
1	ID			q7	q14a	q14b	q14c
2	432		4	7	1	1	
3	434	2006	4	7	1	1	
4	441	2005	6	7	1	2	

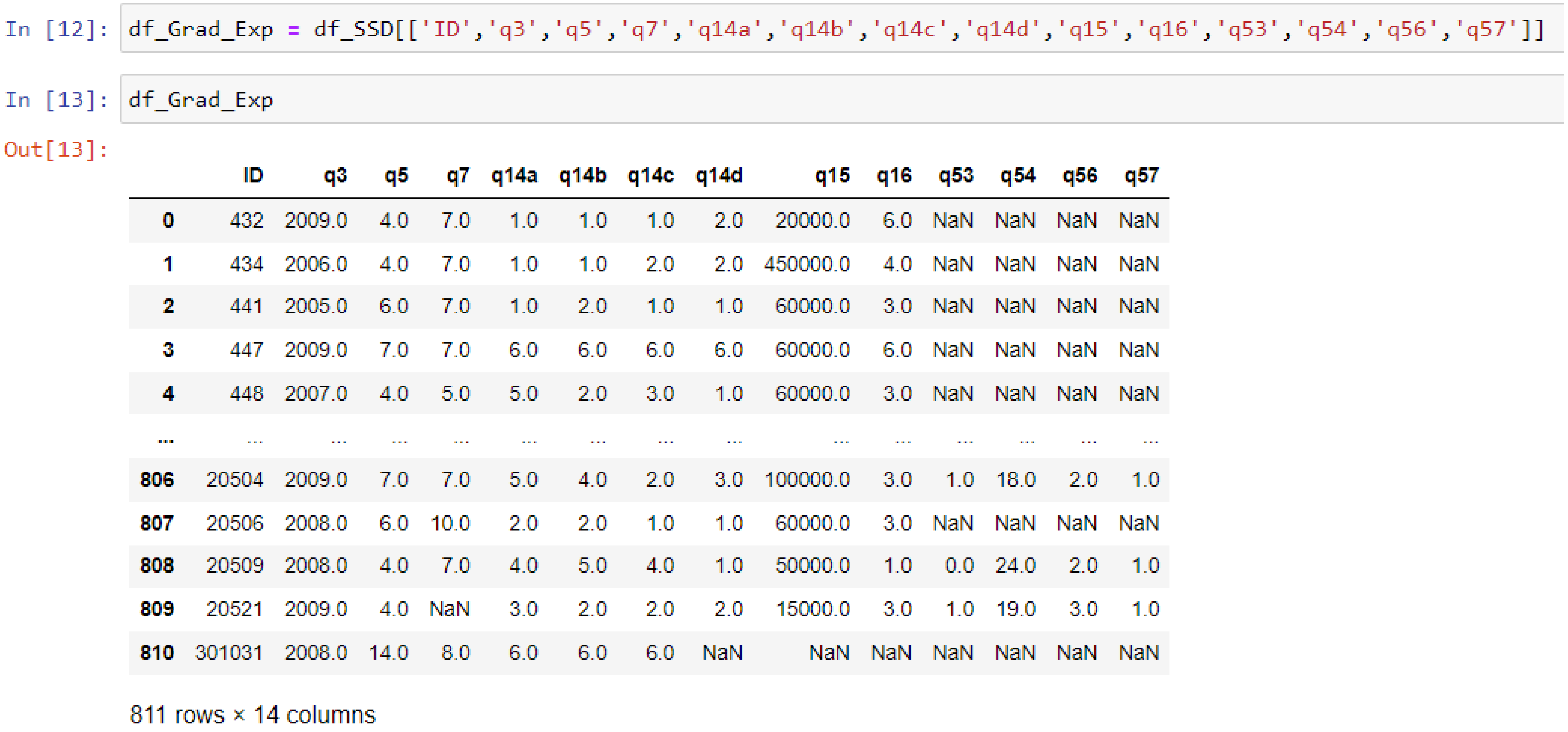
- Excel will normally identify the dimensions of the data file.
- In this case the data file goes from A1 to N812.
- That is, it starts in the top left at column A and row 1, and goes to column N and row 812.
- Excel puts $ signs everywhere $A$1: $N$812. Don’t worry.



	A	B	C	D	E	F	G
1	ID	q3	q5	q7	q14a	q14b	q14c
2	432	2009	4	7	1	1	1
3	434	2006	4	7	1	1	2
4	441	2005	6	7	1	2	1
5	447	2009	7	7	6	6	6

Python

In Python, a Table is called a **dataframe**.



```
In [12]: df_Grad_Exp = df_SSD[['ID','q3','q5','q7','q14a','q14b','q14c','q14d','q15','q16','q53','q54','q56','q57']]

In [13]: df_Grad_Exp

Out[13]:
         ID      q3    q5    q7  q14a  q14b  q14c  q14d       q15  q16  q53  q54  q56  q57
  0     432  2009.0   4.0   7.0   1.0   1.0   1.0   2.0   20000.0  6.0  NaN  NaN  NaN  NaN
  1     434  2006.0   4.0   7.0   1.0   1.0   2.0   2.0  450000.0  4.0  NaN  NaN  NaN  NaN
  2     441  2005.0   6.0   7.0   1.0   2.0   1.0   1.0   60000.0  3.0  NaN  NaN  NaN  NaN
  3     447  2009.0   7.0   7.0   6.0   6.0   6.0   6.0   60000.0  6.0  NaN  NaN  NaN  NaN
  4     448  2007.0   4.0   5.0   5.0   2.0   3.0   1.0   60000.0  3.0  NaN  NaN  NaN  NaN
...     ...     ...   ...   ...   ...   ...   ...   ...       ...  ...  ...  ...  ...  ...
806   20504  2009.0   7.0   7.0   5.0   4.0   2.0   3.0  100000.0  3.0  1.0 18.0  2.0  1.0
807   20506  2008.0   6.0  10.0   2.0   2.0   1.0   1.0   60000.0  3.0  NaN  NaN  NaN  NaN
808   20509  2008.0   4.0   7.0   4.0   5.0   4.0   1.0   50000.0  1.0  0.0 24.0  2.0  1.0
809   20521  2009.0   4.0   NaN   3.0   2.0   2.0   2.0   15000.0  3.0  1.0 19.0  3.0  1.0
810  301031  2008.0  14.0   8.0   6.0   6.0   6.0   NaN       NaN  NaN  NaN  NaN  NaN  NaN

811 rows × 14 columns
```

- Pretty. So what?
- Alternating shades makes it easier to read across rows.
- Each of the headings now has a down arrow. This is powerful. We will look at it next class.

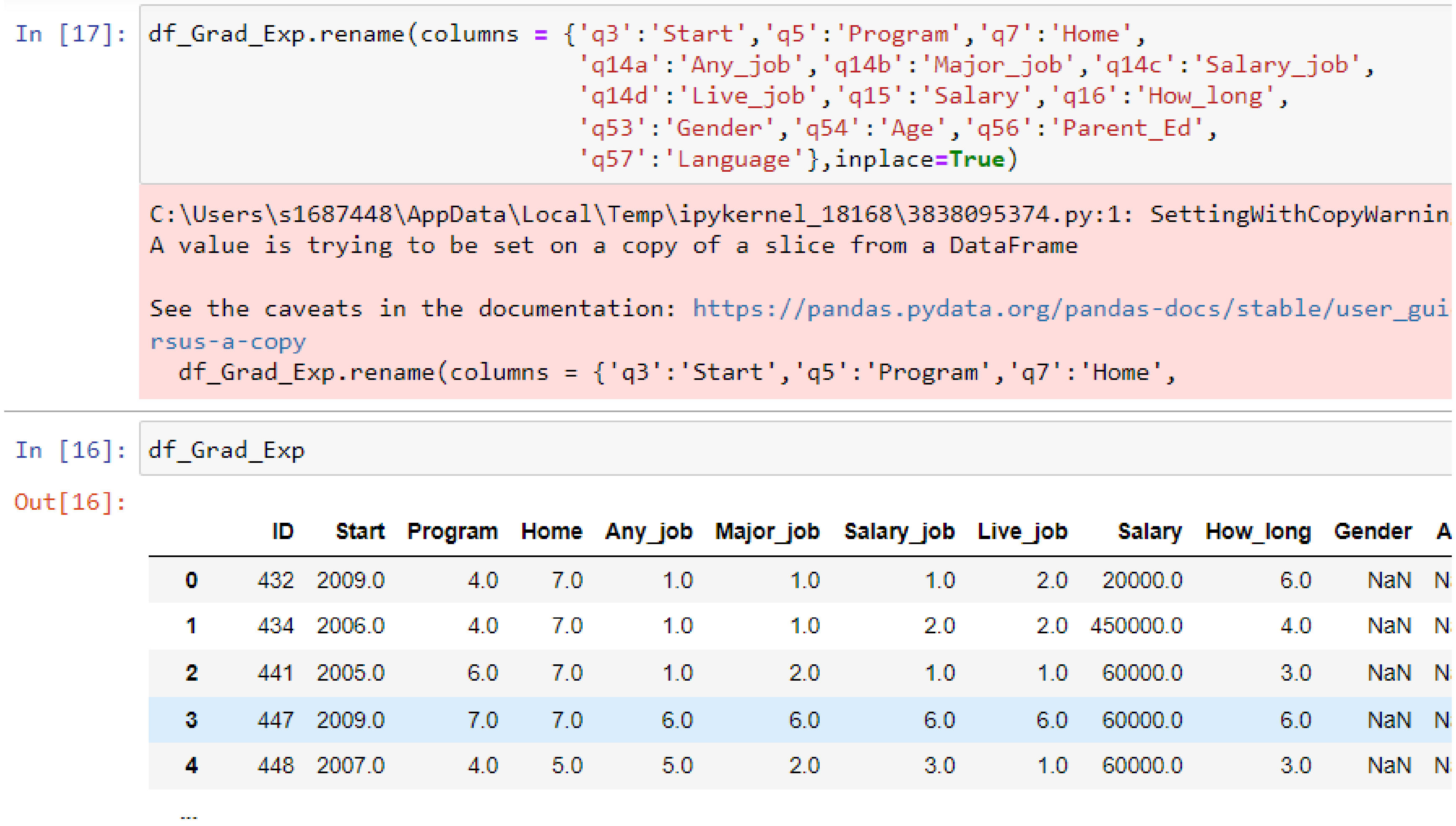
- We have started to make the data set easier to use by selecting only the variables we think important (**Feature Selection**).

- Let us now rename our variables to make then easier to remember.
- Simply retype the labels in the top row with the new names.

q3	start
q5	program
q7	home
q14a	Any job
q14b	Major job
q14c	Salary job
q14d	Live job
q15	Salary
q16	How long
q53	Gender
q54	Age
q56	Parent Ed
q57	Language

Python

In Python, we need to **rename** our variables.



```
In [17]: df_Grad_Exp.rename(columns = {'q3':'Start','q5':'Program','q7':'Home',
                                       'q14a':'Any_job','q14b':'Major_job','q14c':'Salary_job',
                                       'q14d':'Live_job','q15':'Salary','q16':'How_long',
                                       'q53':'Gender','q54':'Age','q56':'Parent_Ed',
                                       'q57':'Language'},inplace=True)
```

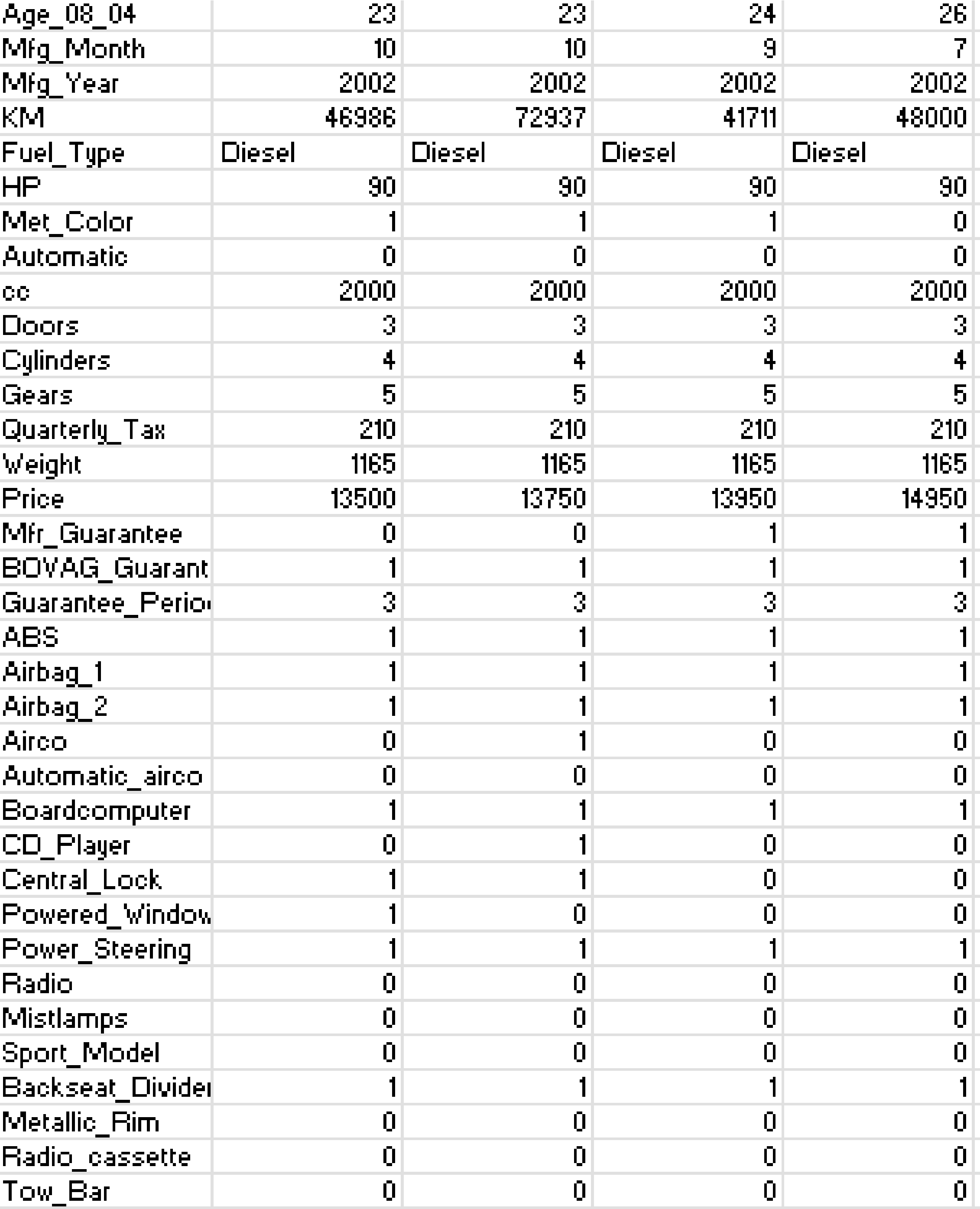
```
C:\Users\s1687448\AppData\Local\Temp\ipykernel_18168\3838095374.py:1: SettingWithCopyWarnin
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gui
rsus-a-copy
  df_Grad_Exp.rename(columns = {'q3':'Start','q5':'Program','q7':'Home',
```

```
In [16]: df_Grad_Exp
```

Out[16]:

	ID	Start	Program	Home	Any_job	Major_job	Salary_job	Live_job	Salary	How_long	Gender	A
0	432	2009.0	4.0	7.0	1.0	1.0	1.0	2.0	20000.0	6.0	NaN	N
1	434	2006.0	4.0	7.0	1.0	1.0	2.0	2.0	450000.0	4.0	NaN	N
2	441	2005.0	6.0	7.0	1.0	2.0	1.0	1.0	60000.0	3.0	NaN	N
3	447	2009.0	7.0	7.0	6.0	6.0	6.0	6.0	60000.0	6.0	NaN	N
4	448	2007.0	4.0	5.0	5.0	2.0	3.0	1.0	60000.0	3.0	NaN	N

...



Age_08_04	23	23	24	26
Mfg_Month	10	10	9	7
Mfg_Year	2002	2002	2002	2002
KM	46986	72937	41711	48000
Fuel_Type	Diesel	Diesel	Diesel	Diesel
HP	90	90	90	90
Met_Color	1	1	1	0
Automatic	0	0	0	0
cc	2000	2000	2000	2000
Doors	3	3	3	3
Cylinders	4	4	4	4
Gears	5	5	5	5
Quarterly_Tax	210	210	210	210
Weight	1165	1165	1165	1165
Price	13500	13750	13950	14950
Mfr_Guarantee	0	0	1	1
BOVAG_Guarant	1	1	1	1
Guarantee_Perio	3	3	3	3
ABS	1	1	1	1
Airbag_1	1	1	1	1
Airbag_2	1	1	1	1
Airco	0	1	0	0
Automatic_airco	0	0	0	0
Boardcomputer	1	1	1	1
CD_Player	0	1	0	0
Central_Lock	1	1	0	0
Powered_Window	1	0	0	0
Power_Steering	1	1	1	1
Radio	0	0	0	0
Mistlamps	0	0	0	0
Sport_Model	0	0	0	0
Backseat_Divider	1	1	1	1
Metallic_Rim	0	0	0	0
Radio_cassette	0	0	0	0
Tow_Bar	0	0	0	0

This data set was downloaded from Kaggle. It is said to represent used Toyota Corollas sold in the Netherlands in 2004.

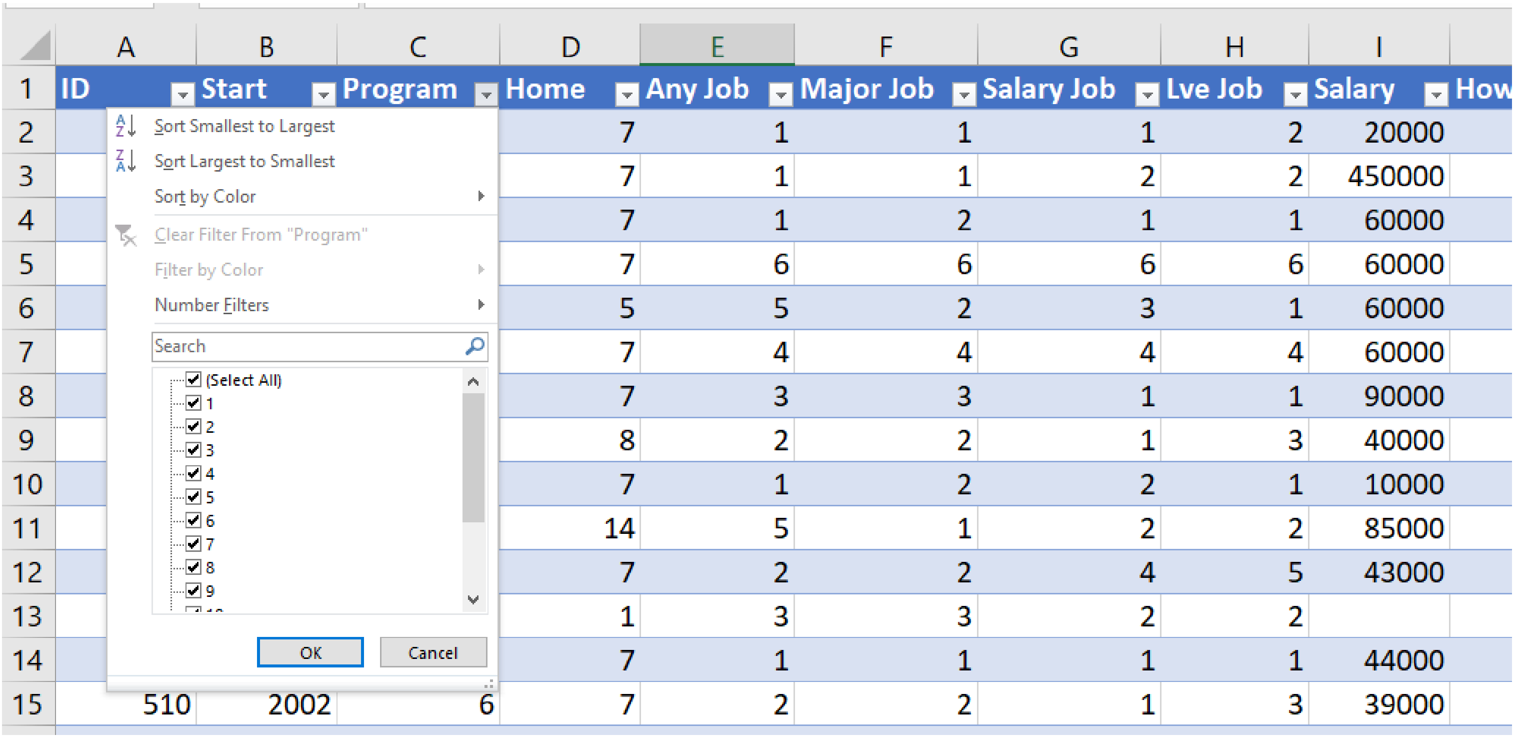
Like most data files on Kaggle, no data dictionary was provided.

Although descriptive names are given to variables, some are still ambiguous.

Many company databases have 100s or 1000s of fields with short names. They can be confusing.

- Some of the data values are difficult to understand.
- Let us transform them into more useable values.
- If you click on the down arrow by a variable, you will get a list of values, and things you can do.

- Select **Program**.



	A	B	C	D	E	F	G	H	I	
1	ID	Start	Program	Home	Any Job	Major Job	Salary Job	Lve Job	Salary	How
2				7	1	1	1	2	20000	
3				7	1	1	2	2	450000	
4				7	1	2	1	1	60000	
5				7	6	6	6	6	60000	
6				5	5	2	3	1	60000	
7				7	4	4	4	4	60000	
8				7	3	3	1	1	90000	
9				8	2	2	1	3	40000	
10				7	1	2	2	1	10000	
11				14	5	1	2	2	85000	
12				7	2	2	4	5	43000	
13				1	3	3	2	2		
14				7	1	1	1	1	44000	
15	510	2002	6	7	2	2	1	3	39000	

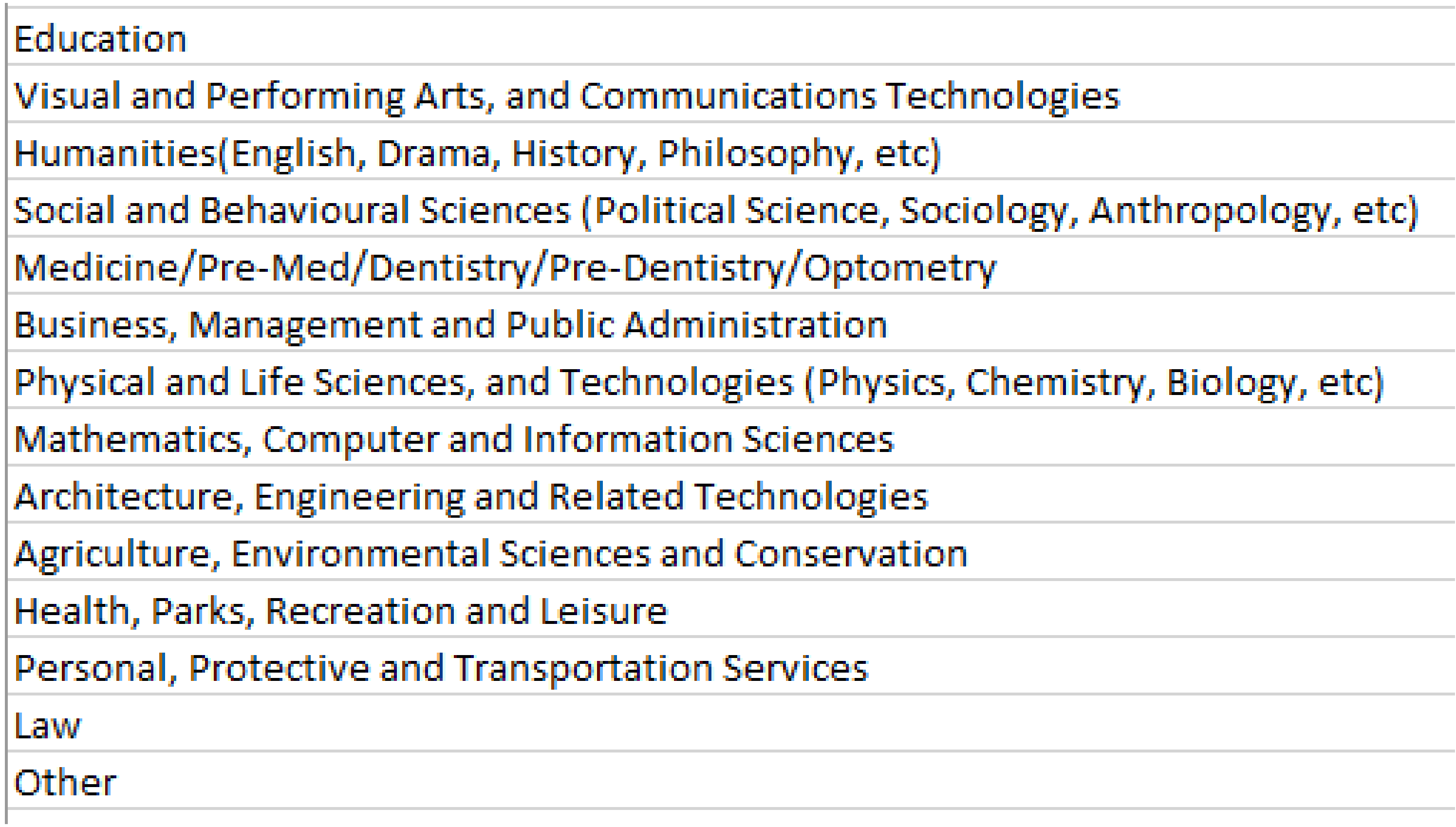
- How do we interpret the various values that are there?
- What does it mean to have a Program = 4?
- From the Interpreted Data tab we know that the 4th entry is Social Sciences.
- Rather than having to memorize the meanings, can we create a new variables with descriptive values?

Transforming Data

- Insert a new column to the left of **Program** and call it **Program**.
- Excel will rename the original Program column, **Program2**. In a Table, each column must have a unique name. No 2 columns can have the same name.

- We need to translate the numeric codes for Program2 to the verbal descriptions we want in Program.

- Create a new worksheet and call it **Lookup Tables**.
- A look up table is simply a table that matches one set of values to a new set of values. We have programs numbered 1 to 14 that we want to match to new descriptions.

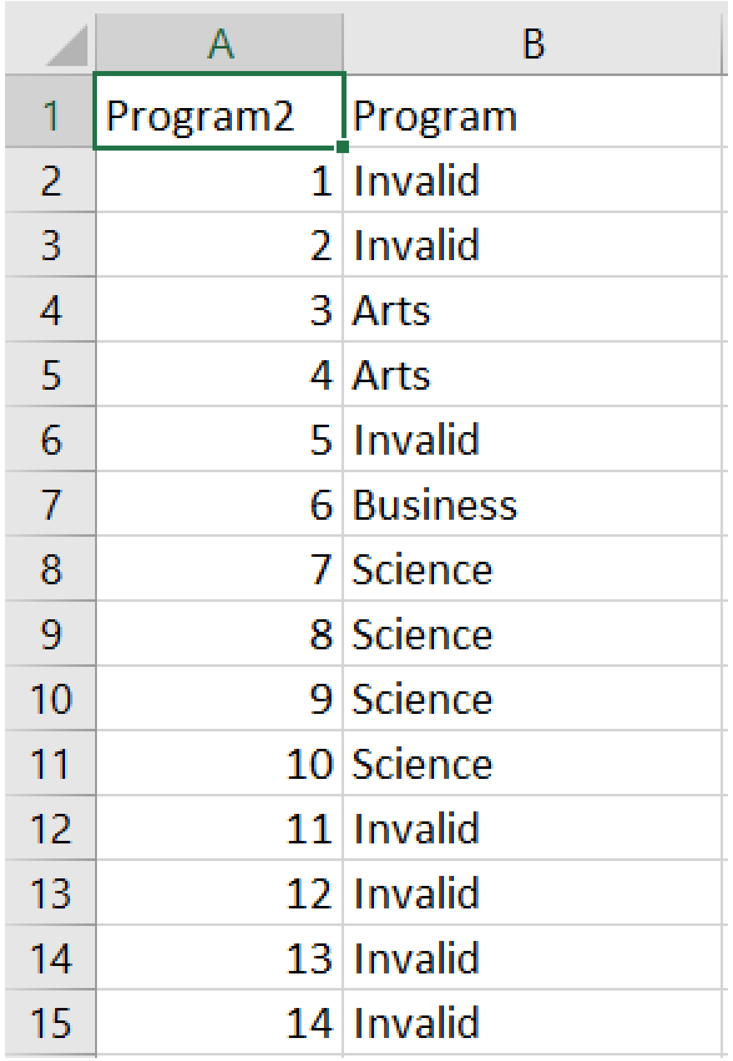


Education
Visual and Performing Arts, and Communications Technologies
Humanities(English, Drama, History, Philosophy, etc)
Social and Behavioural Sciences (Political Science, Sociology, Anthropology, etc)
Medicine/Pre-Med/Dentistry/Pre-Dentistry/Optometry
Business, Management and Public Administration
Physical and Life Sciences, and Technologies (Physics, Chemistry, Biology, etc)
Mathematics, Computer and Information Sciences
Architecture, Engineering and Related Technologies
Agriculture, Environmental Sciences and Conservation
Health, Parks, Recreation and Leisure
Personal, Protective and Transportation Services
Law
Other

- This was a national survey and the list was meant to reflect the wide range of programs that a student may be enrolled in.
- But we do not offer many of these programs.
- Students selected their program from this list.
- Why would a student at Saint Mary’s select Education?
- Or Medicine?
- Or Protective Services?
- Some Saint Mary’s students made these choices.

- We only offer Arts, Business, Science and Engineering.
- What should we do?
- Guess what the right choice should be?
- Exclude them? Let us mark them as Invalid.

In your Lookup Tables sheet, create a table like the one below.

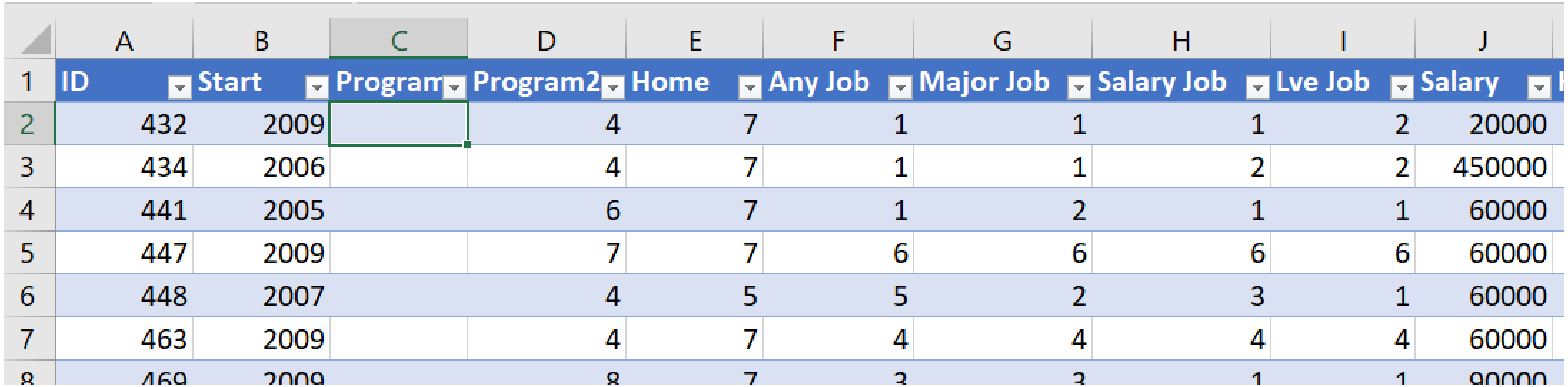


	A	B
1	Program2	Program
2	1	Invalid
3	2	Invalid
4	3	Arts
5	4	Arts
6	5	Invalid
7	6	Business
8	7	Science
9	8	Science
10	9	Science
11	10	Science
12	11	Invalid
13	12	Invalid
14	13	Invalid
15	14	Invalid

- It is desirable to keep all data in one sheet and lookup tables in a separate sheet.
- As you do recoding and then analysis, if all done in the same sheet, you get a lot of clutter.

- Excel has a function called **VLOOKUP**.
- It uses a “look up” table, with the original value in one column and the new value in the 2nd column.

- Return to the **Grad Expectations** sheet.



	A	B	C	D	E	F	G	H	I	J
1	ID	Start	Program	Program2	Home	Any Job	Major Job	Salary Job	Lve Job	Salary
2	432	2009		4	7	1	1	1	2	20000
3	434	2006		4	7	1	1	2	2	450000
4	441	2005		6	7	1	2	1	1	60000
5	447	2009		7	7	6	6	6	6	60000
6	448	2007		4	5	5	2	3	1	60000
7	463	2009		4	7	4	4	4	4	60000
8	469	2009		8	7	3	3	1	1	90000

In the cell C2, type the following

=VLOOKUP(D2,

You want Excel to take the value in D2 and then assign C2 the new value based upon the Look Up table.

- Click on the Look Up table sheet and select the whole table, A2:B9. Do not include the header row.

The Look Up table will always be between A2 and B15, so put $ everywhere to fix this location.

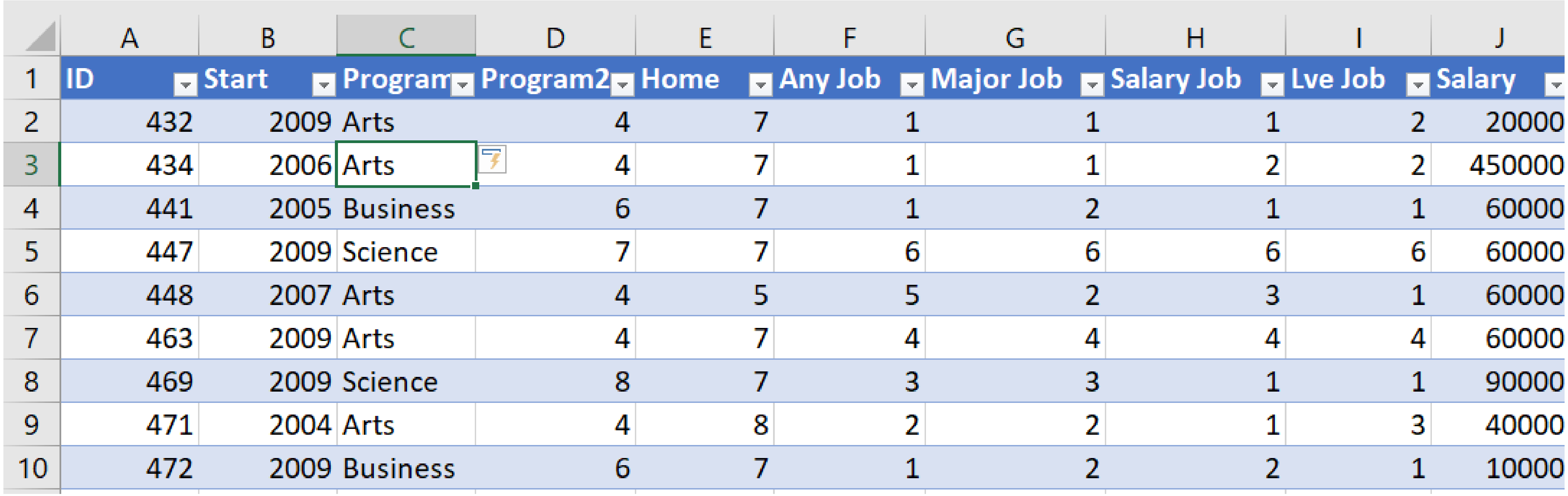
=VLOOKUP(D2,'look up'!$A$2:$B$15,

The new value we wish to assign is in the second column of the table, so enter 2.

We want an **Exact** match of value in column A to those in column B, so enter **False**.

=VLOOKUP(C2,'look up'!$A$2:$B$15,2,FALSE)

- Observe that Excel fills in all the entries in column C
- With a Table, Excel assumes that the formula applies to all rows.



	A	B	C	D	E	F	G	H	I	J
1	ID	Start	Program	Program2	Home	Any Job	Major Job	Salary Job	Lve Job	Salary
2	432	2009	Arts	4	7	1	1	1	2	20000
3	434	2006	Arts	4	7	1	1	2	2	450000
4	441	2005	Business	6	7	1	2	1	1	60000
5	447	2009	Science	7	7	6	6	6	6	60000
6	448	2007	Arts	4	5	5	2	3	1	60000
7	463	2009	Arts	4	7	4	4	4	4	60000
8	469	2009	Science	8	7	3	3	1	1	90000
9	471	2004	Arts	4	8	2	2	1	3	40000
10	472	2009	Business	6	7	1	2	2	1	10000

- As you start your exploration, look at the data
- Look at **ALL** the data – every variable
- Do the values appear reasonable?

- With cases where individuals do data entry, there can be errors
  - Choose the wrong item from a list
  - Data entry error
  - Misunderstood question
  - Dishonest

- In this case, students may not have understood the choices.

Social and Behavioural Sciences (Political Science, Sociology, Anthropology)

- What about Psychology or Criminology?
- Maybe I want to go into Education or Medicine and I am in a pathway

- Similarly, we can recode Home, Gender, Parent Ed and Language.
- I have done this and saved the recoded table in a new tab called **Grad Recoded**.

- Look at what happened to Gender, Parent Ed and Language.
- For Parent Ed and Language, the entries are #NA. Why?

- For Gender, we have blanks that appear as Male. Why?

- Gender was one of the last questions on the survey.
- Many students did not complete the survey (fatigue) and left the question blank.
- It is unlikely that they were embarrassed to select gender.
- In 2010, it was not common to identify publicly as non-binary.
- In the survey, Male/Female had been coded as 0/1.
- Excel will treat a blank as 0, and thus treat blank as Male.

- When recoding variables, you may want to use an IF statement to treat blanks differently.
- If 0 is not in your lookup table, Excel will recode blanks as #NA (not applicable).

For Gender, rather than use

=VLOOKUP(O2,'look up'!$K$2:$L$3,2,FALSE)

Type

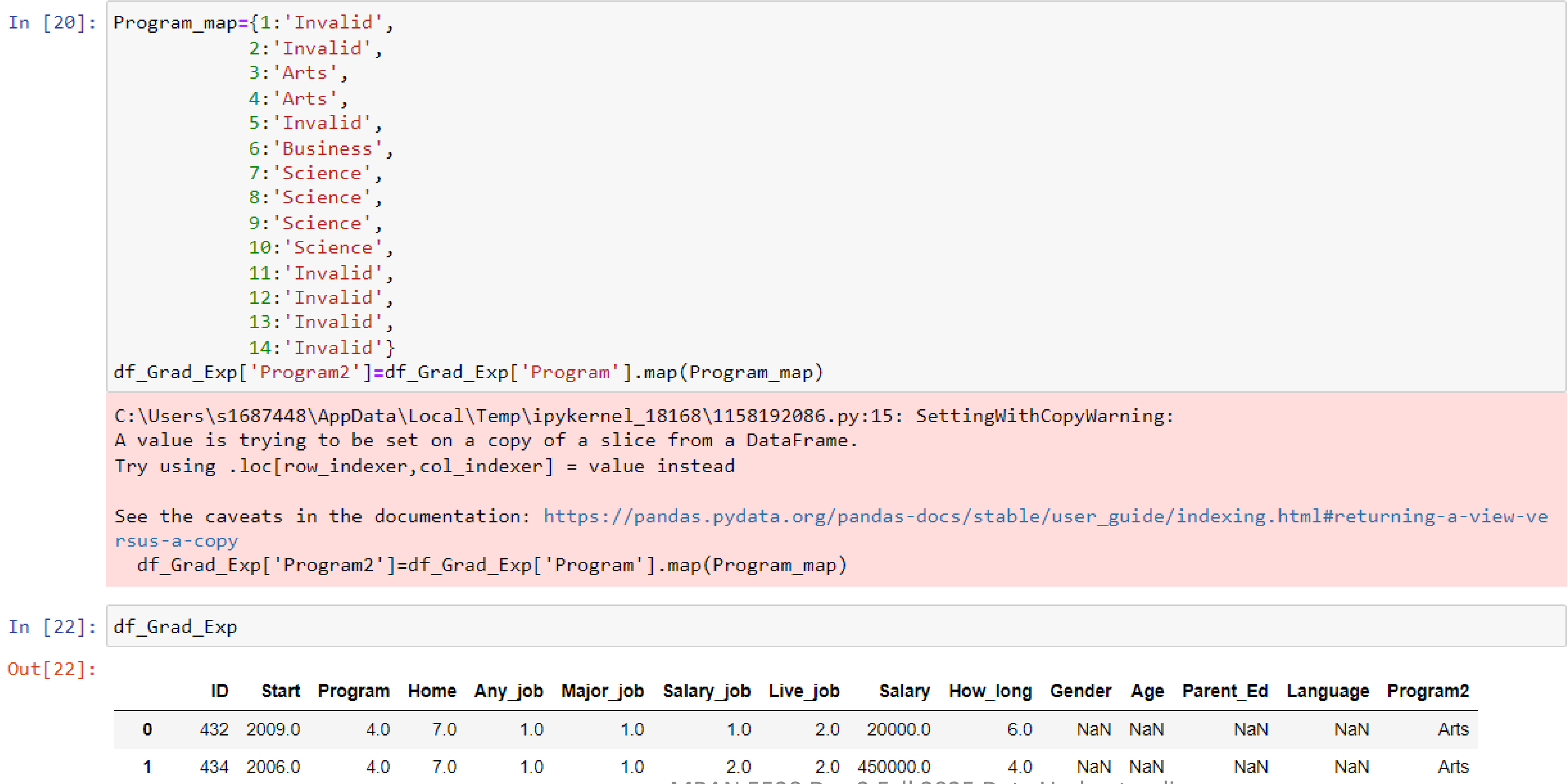
=IF(O2=“”,””,VLOOKUP(O2,'look up'!$K$2:$L$3,2,FALSE))

This says,

- if O2 is blank,
  - leave it blank,
- otherwise,
  - use VLOOKUP

Python

In Python, there are multiple ways to recode values. I used a **dict** to **map** values, but you can also do it by using a **join**. I am not good with joins.



```
In [20]: Program_map={1:'Invalid',
                      2:'Invalid',
                      3:'Arts',
                      4:'Arts',
                      5:'Invalid',
                      6:'Business',
                      7:'Science',
                      8:'Science',
                      9:'Science',
                      10:'Science',
                      11:'Invalid',
                      12:'Invalid',
                      13:'Invalid',
                      14:'Invalid'}
         df_Grad_Exp['Program2']=df_Grad_Exp['Program'].map(Program_map)
```

```
C:\Users\s1687448\AppData\Local\Temp\ipykernel_18168\1158192086.py:15: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
  df_Grad_Exp['Program2']=df_Grad_Exp['Program'].map(Program_map)
```

```
In [22]: df_Grad_Exp
```

Out[22]:

	ID	Start	Program	Home	Any_job	Major_job	Salary_job	Live_job	Salary	How_long	Gender	Age	Parent_Ed	Language	Program2
0	432	2009.0	4.0	7.0	1.0	1.0	1.0	2.0	20000.0	6.0	NaN	NaN	NaN	NaN	Arts
1	434	2006.0	4.0	7.0	1.0	1.0	2.0	2.0	450000.0	4.0	NaN	NaN	NaN	NaN	Arts

What is an **Approximate Match**?

Suppose that with the Salaries, we don’t need to know the exact amount but simply whether the expected salary is very low, low, average, high, very high, or unrealistic?

Under $20,000	= Very Low
$20,000 to $40,000	= Low
$40,000 to $60,000	= Average
$60,000 to $80,000	= High
$80,000 to $100,000	= Very High
Above $120,000	= Unrealistic

Looks like a Look Up table with groups of values.

Excel only needs to know the **starting** value for each group.
The starting value for one group is the ending value for the previous group.

Salary	Salary Grp
0	Very Low
20000	Low
40000	Average
60000	High
80000	Very High
100000	Unrealistic

- Create a lookup table in the Lookup Table sheet for the above.
- I put mine in O1:P7

- Create a new column to the left of **Salary** and call it **Salary Grp**.

- In cell **K2**, type
- **=VLOOKUP(L2,'look up’!$O$2:$P$7,2,TRUE)**

- By selecting **TRUE** instead of FALSE, Excel looks for an Approximate Match.
- Each value is assigned to the highest group that it can fit.
- For example, $50,000 is greater than $40,000 but not as large as $60,000 so it is in the Average group.

- Explore the recoded data file.
- The values should now be easier to interpret.

- For example, we can now see that the entry in row 10, ID 472, is
  - A first year student (started in 2009) who is studying Business (program2=6) and is from Nova Scotia.
  - She is female, parents’ education is unknown and speaks English at home.
  - Her salary expectations are very low.

- We have covered a lot!

- In Excel, we have:
  - Created a new file to protect the original data.
  - Used Table to make our data a special data table.
  - Renamed variables to make them easier to remember.
  - Recoded values using VLOOKUP with an exact match.
  - Used an IF statement to ensure that blanks and zeros are recorded correctly.
  - Started aggregating data with an approximate match.

- From a Data Science perspective, we have
  - Seen the need for a data dictionary so we know what the data means.
  - Noted the importance of protecting data.
  - Seen the value in “preparing” data for analysis.
  - Seen that recoding data may have unintended consequences.
  - Observed data quality issues: missing values, data entry errors, misunderstanding questions, …

Data Preparation best practices

1. Always protect the original data from damage. Select only the variables you think you may need and put them in your working data file.
2. Make the data talk to you in language you understand:
   a. Give variables meaningful names.
   b. Recode values with understandable values.
   c. Remove unnecessary detail by grouping values if appropriate.
   d. Make note of anything unusual. Such cases should be investigated.
   e. Keep track of what you have done and keep these notes with your working data file (e.g., keep a data dictionary, keep all your look up tables used for recoding, …). It is easy to forget what changes you have made.

What is next?

- Explore characteristics of a single variable.
- What are typical values? Where is the middle of the data set?
- What is a low(high) value?
- What is unusually low (high)?
- What would we consider an outlier?
- Can we summarize the overall pattern of the data? Numerically/Graphically
- What are common patterns?

- Although we are moving into data exploration and thinking about modeling, we may also have to think more about data preparation (e.g., dealing with outliers, missing values,…)

- What are tools/strategies for exploring relationships among variables?
- What strategies are applicable when the variables are numeric?
- What if one is numeric but others are categorical? (e.g., salary by program)
- What if two variables are categorical? (e.g., concern about job by gender)

Nepal courtesy of Mandipa