Data Analytics and Visualizations with Microsoft Excel 2019

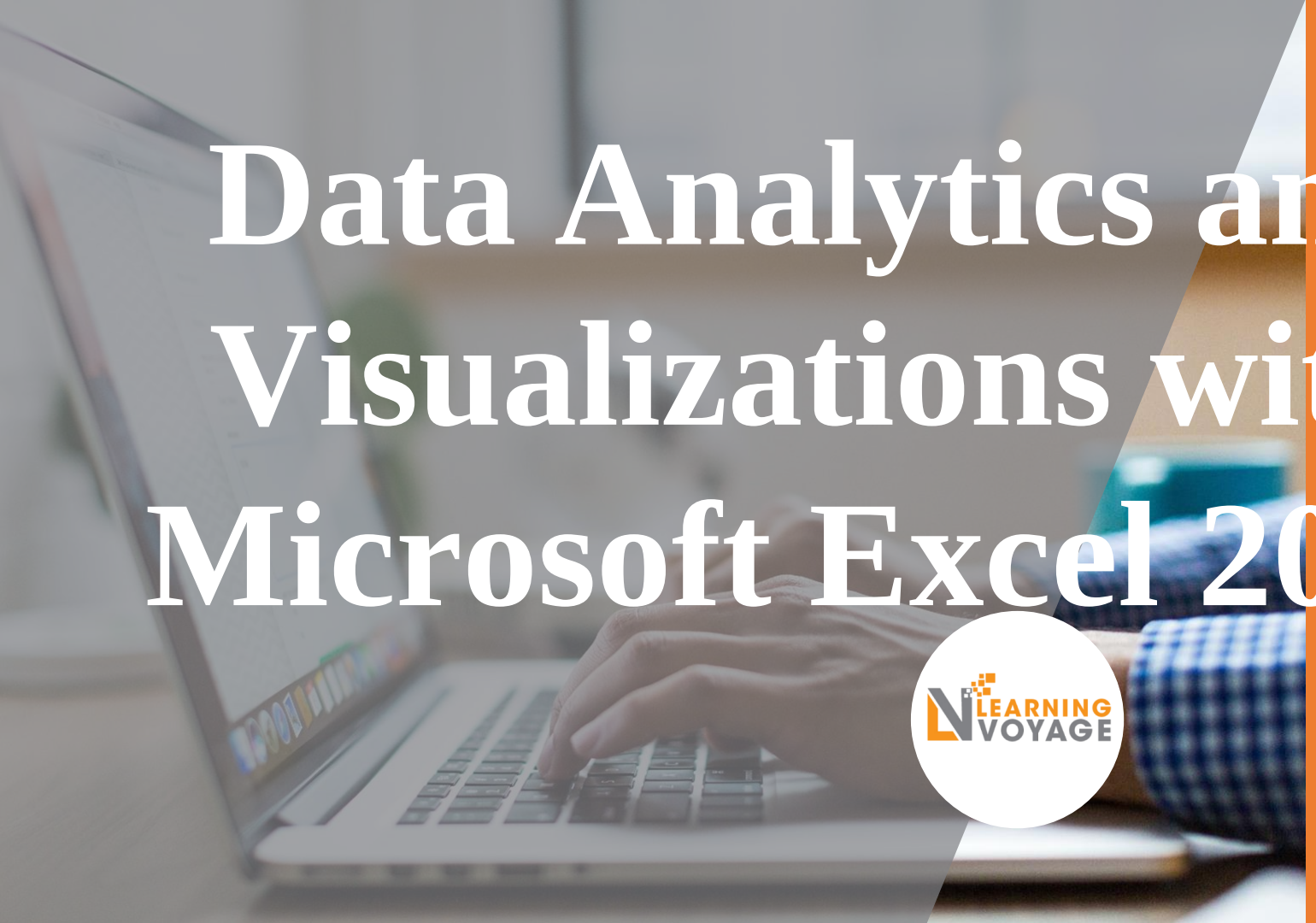
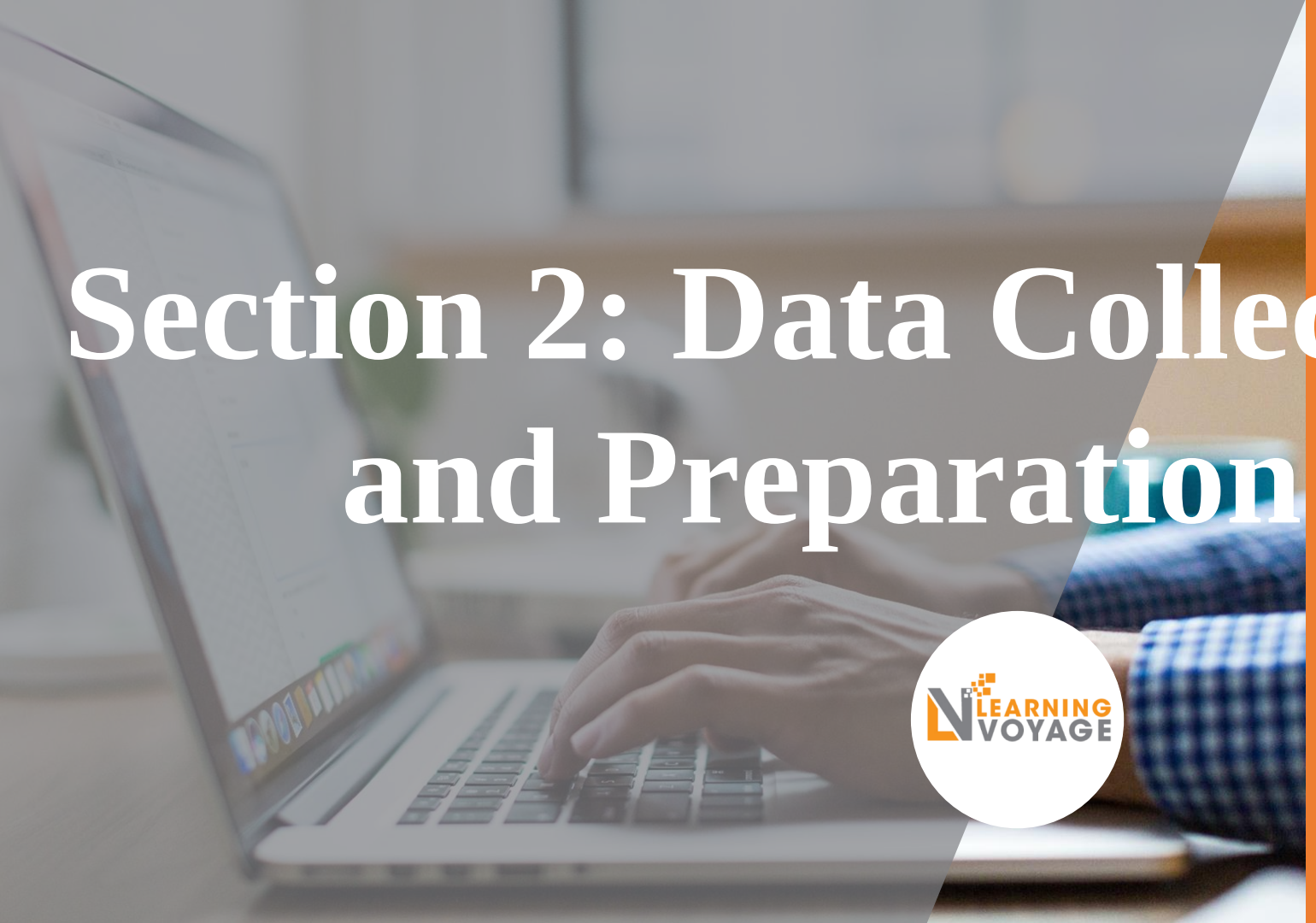


Table of Contents

Table of Contents

Section 2: Data Collection and Preparation

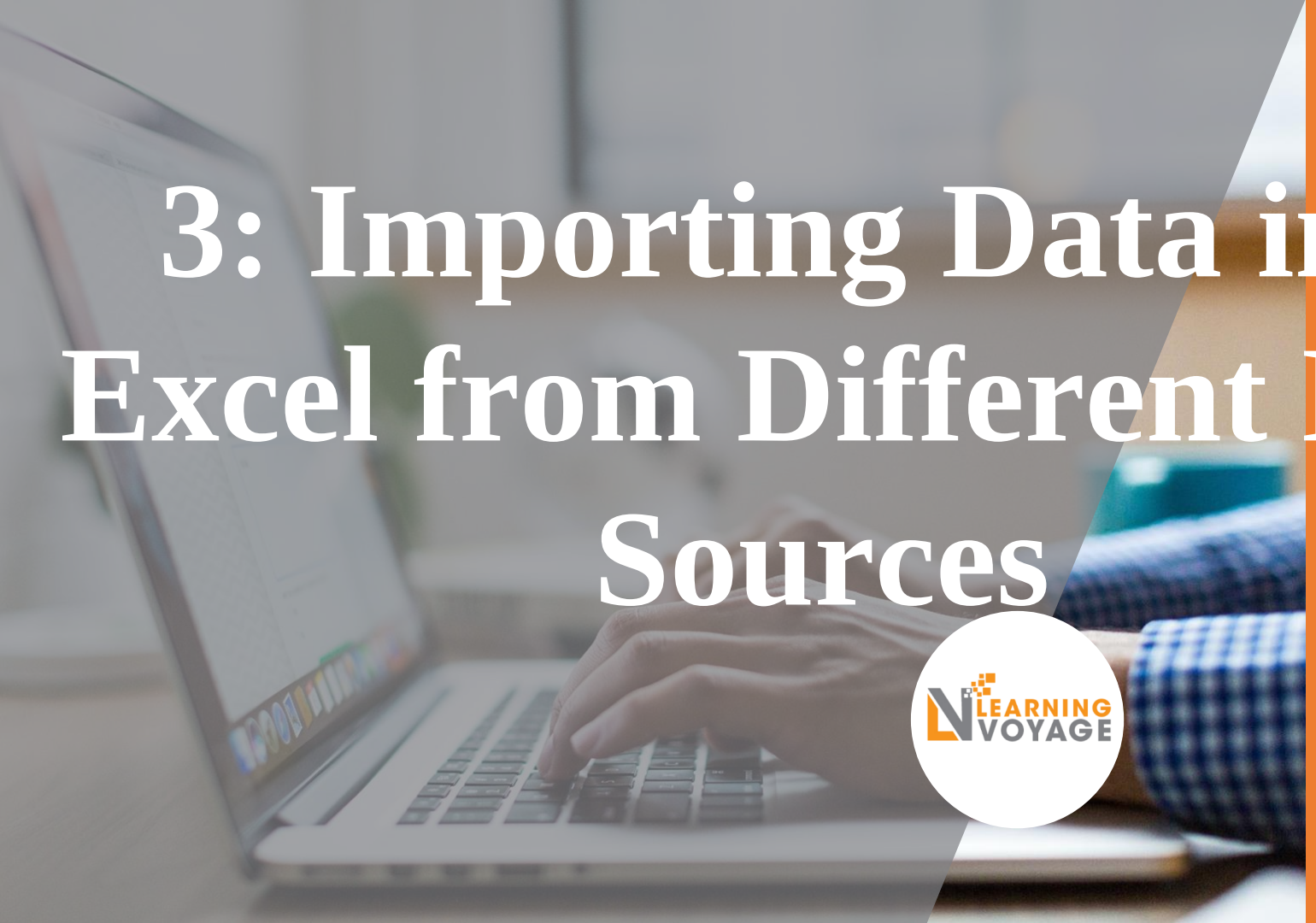


Data Collection and Preparation

This section comprises the following lessons:

- lesson 3, Importing Data into Excel from Different Data Sources
- lesson 4, Data Cleansing and Preliminary Data Analysis
- lesson 5, Correlations and the Importance of Variables

3: Importing Data into Excel from Different Data Sources



Importing Data into Excel from Different Data Sources

In this lesson, we will cover the following topics:

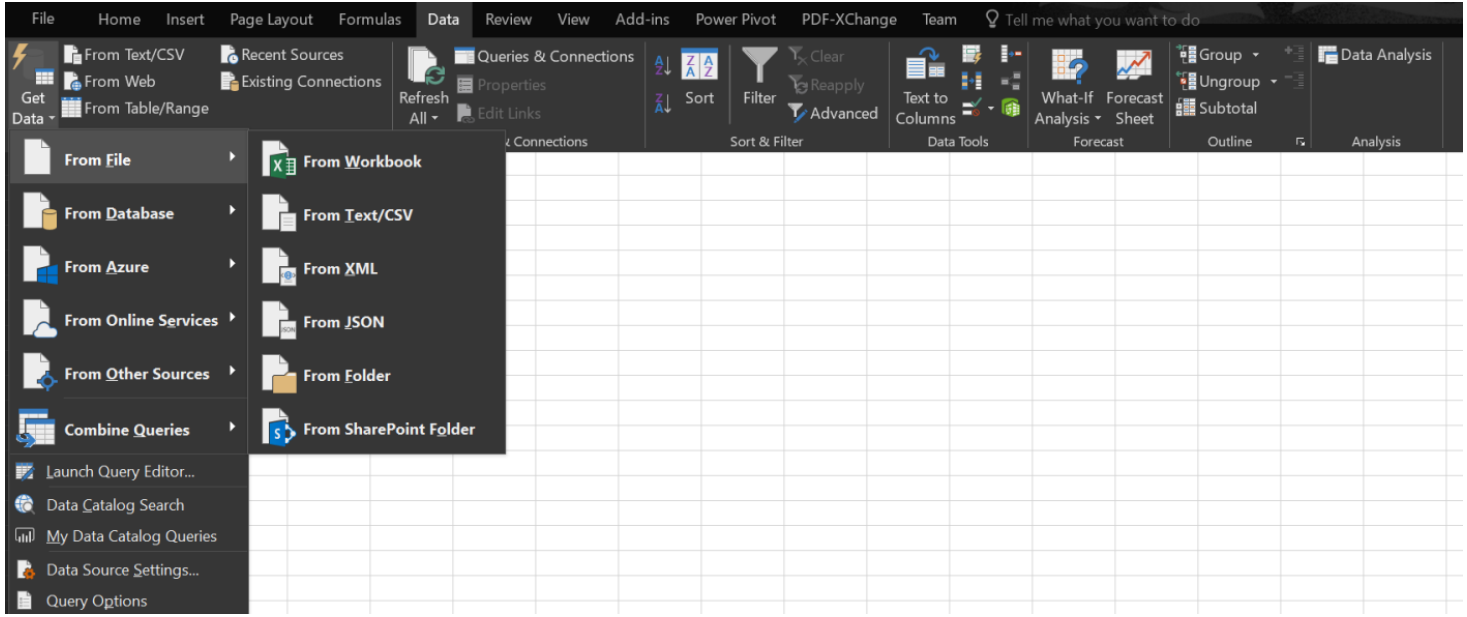
- Importing data from a text file
- Importing data from another Excel workbook
- Importing data from a web page
- Importing data from Facebook
- Importing data from a JSON file
- Importing data from a database

Technical requirements

- You will need to download the homes.csv, homes.txt, titanic.xls, and azure_text_analytics.json files

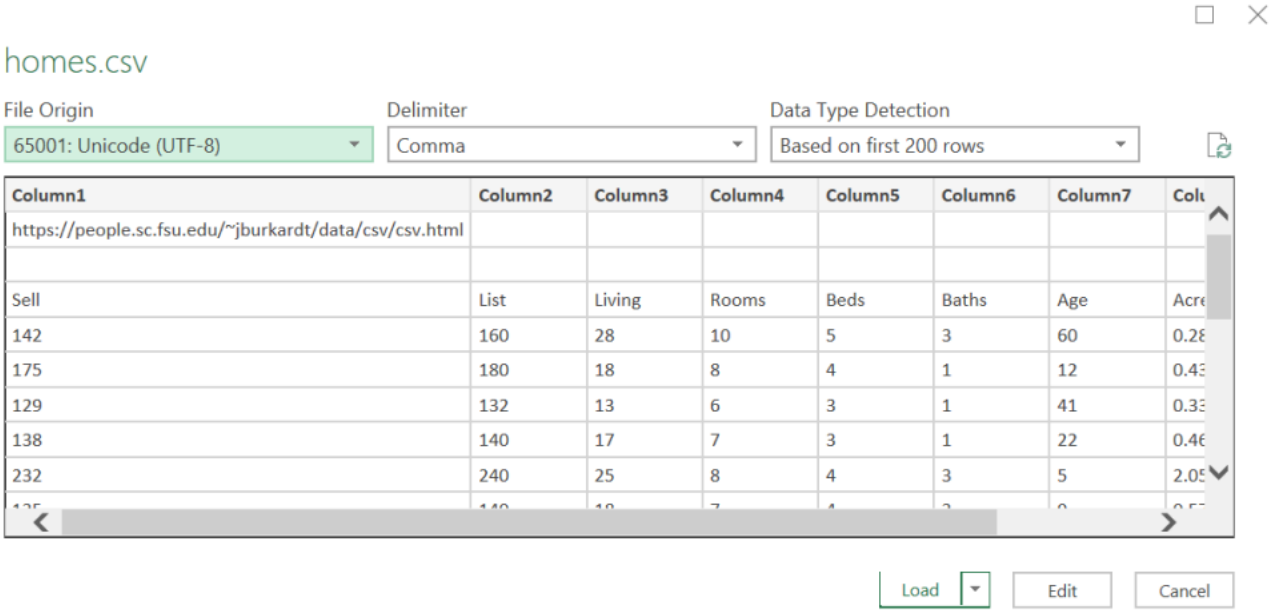
Importing data from a text file

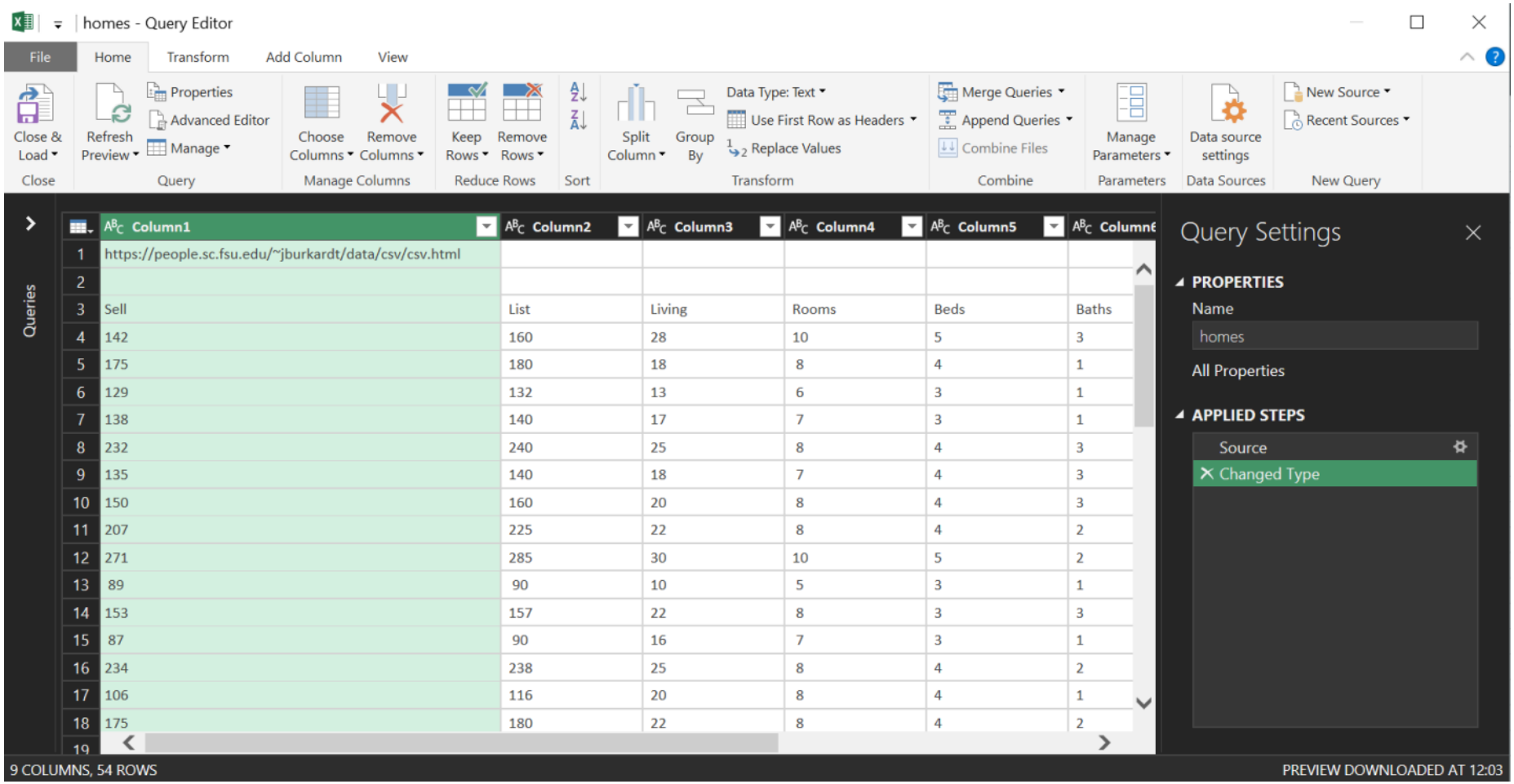
- Click on Data.
- Navigate to Get Data | From File | From Text/CSV:



Importing data from a text file

- A window will pop up, showing you a preview of the file's contents, as shown in the following screenshot:

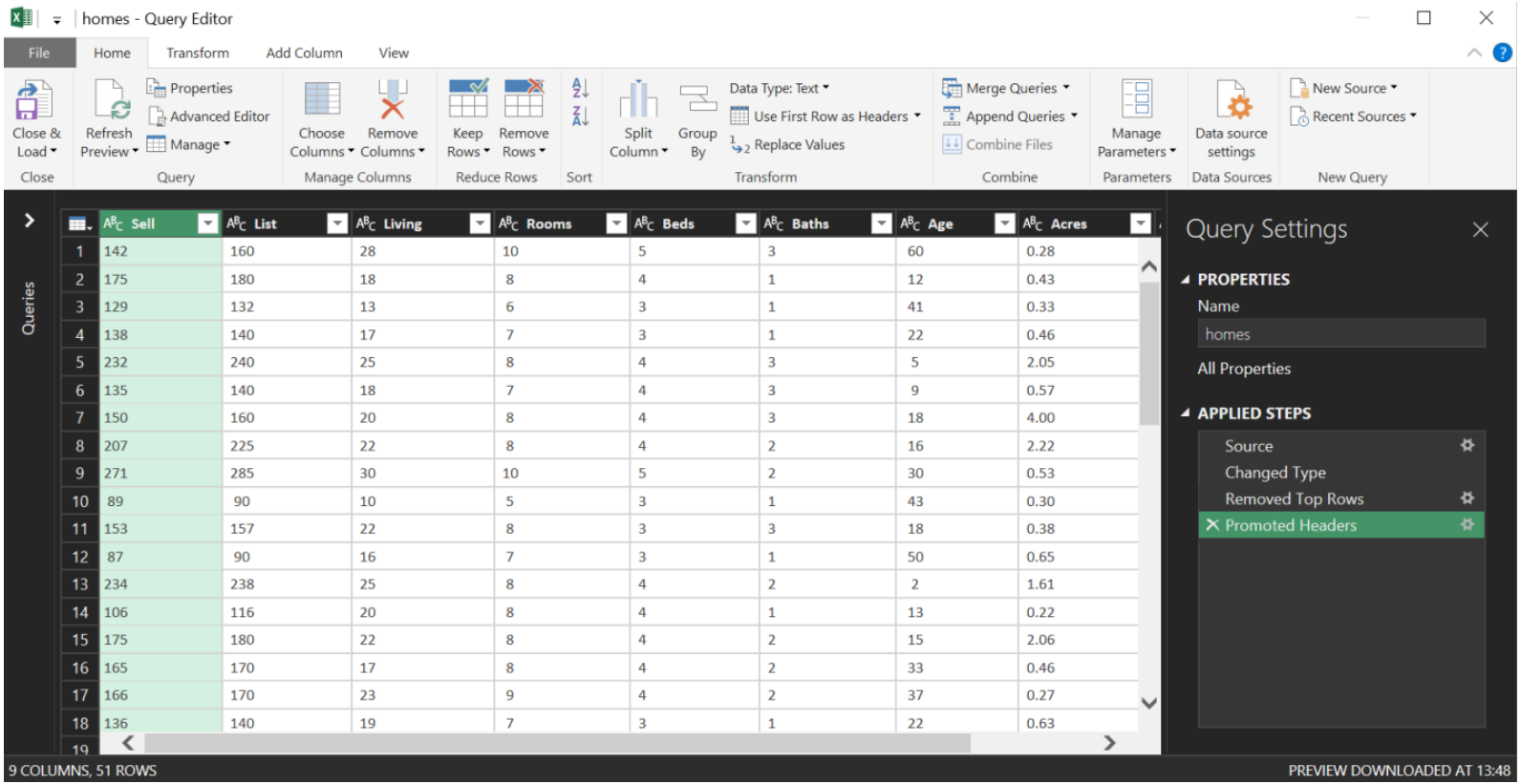



homes - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
Column1
Column2
Column3
Column4
Column5
Column6
https://people.sc.fsu.edu/~jburkardt/data/csv/csv.html
Sell
List
Living
Rooms
Beds
Baths
142 160 28 10 5 3
175 180 18 8 4 1
129 132 13 6 3 1
138 140 17 7 3 1
232 240 25 8 4 3
135 140 18 7 4 3
150 160 20 8 4 3
207 225 22 8 4 2
271 285 30 10 5 2
89 90 10 5 3 1
153 157 22 8 3 3
87 90 16 7 3 1
234 238 25 8 4 2
106 116 20 8 4 1
175 180 22 8 4 2
Query Settings
PROPERTIES
Name
homes
All Properties
APPLIED STEPS
Source
Changed Type
9 COLUMNS, 54 ROWS
PREVIEW DOWNLOADED AT 12:03

Importing data from a text file

- Navigate to Remove Rows | Remove Top Rows.
- You will see the option to specify how many rows you want to skip. In this file, we need to skip 2 rows, as shown in the following screenshot:

- The result of this is shown in the following screenshot:

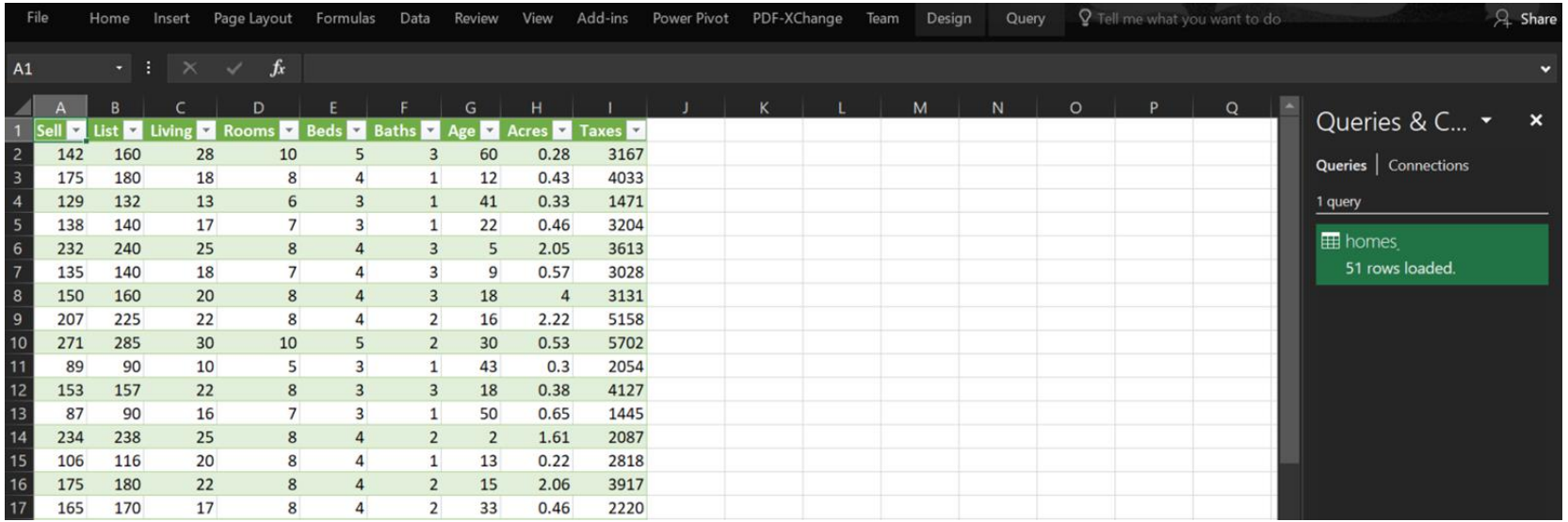


Importing data from a text file

- Select Acres.
- Navigate to Data Type in the Transform menu.
- Change the type to Decimal Number.
- Select the rest of the columns and change the type to Whole Number to fix the other columns.

Importing data from a text file

- Finally, click on Close & Load. You will see the following data table:



	Sell	List	Living	Rooms	Beds	Baths	Age	Acres	Taxes
2	142	160	28	10	5	3	60	0.28	3167
3	175	180	18	8	4	1	12	0.43	4033
4	129	132	13	6	3	1	41	0.33	1471
5	138	140	17	7	3	1	22	0.46	3204
6	232	240	25	8	4	3	5	2.05	3613
7	135	140	18	7	4	3	9	0.57	3028
8	150	160	20	8	4	3	18	4	3131
9	207	225	22	8	4	2	16	2.22	5158
10	271	285	30	10	5	2	30	0.53	5702
11	89	90	10	5	3	1	43	0.3	2054
12	153	157	22	8	3	3	18	0.38	4127
13	87	90	16	7	3	1	50	0.65	1445
14	234	238	25	8	4	2	2	1.61	2087
15	106	116	20	8	4	1	13	0.22	2818
16	175	180	22	8	4	2	15	2.06	3917
17	165	170	17	8	4	2	33	0.46	2220

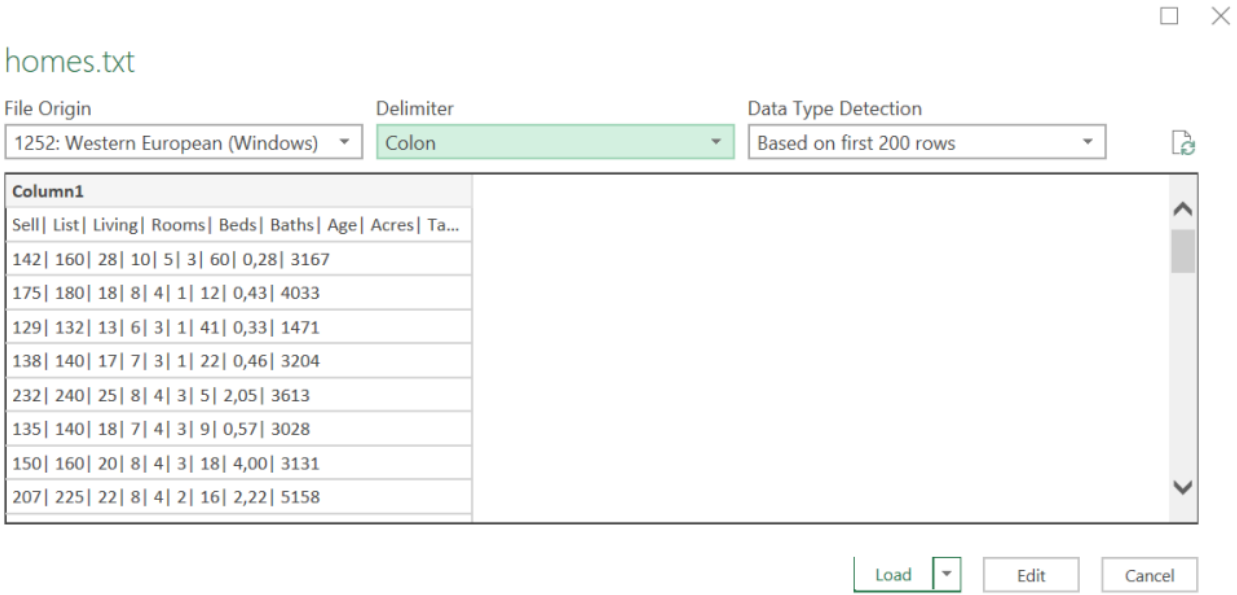
Importing data from a text file

If the file is not a CSV file but just a text file that's using a different separator, we can still load it using a similar procedure. We just repeat the steps we used for importing the CSV file:

- Click on Data.
- Navigate to Get Data | From File | From Text/CSV.

Importing data from a text file

- Navigate to the file's location and open homes.txt. You will see the following preview:



homes.txt

File Origin: 1252: Western European (Windows)

Delimiter: --Custom--

Data Type Detection: Based on first 200 rows

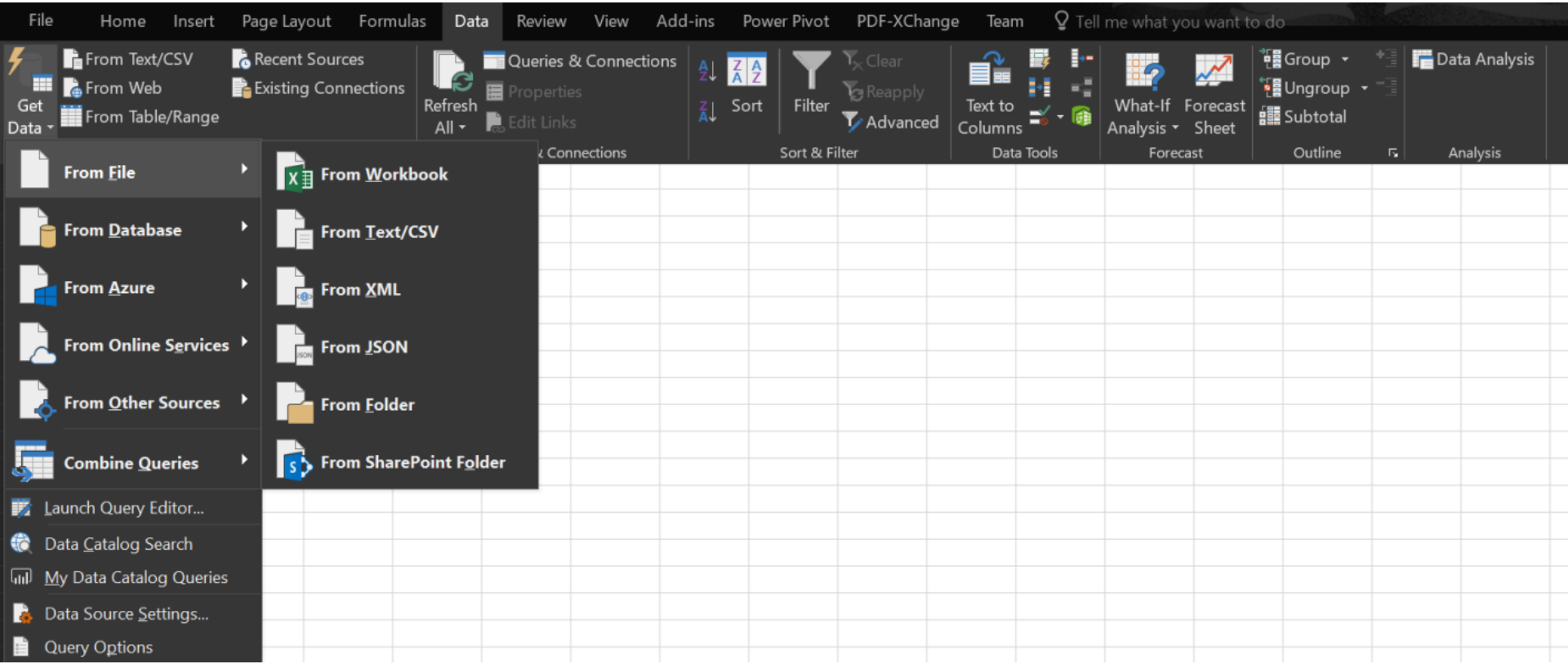
Sell	List	Living	Rooms	Beds	Baths	Age	Acres	Taxes
142	160	28	10	5	3	60	0.28	3167
175	180	18	8	4	1	12	0.43	4033
129	132	13	6	3	1	41	0.33	1471
138	140	17	7	3	1	22	0.46	3204
232	240	25	8	4	3	5	2.05	3613
135	140	18	7	4	3	9	0.57	3028
150	160	20	8	4	3	18	4	3131
207	225	22	8	4	2	16	2.22	5158

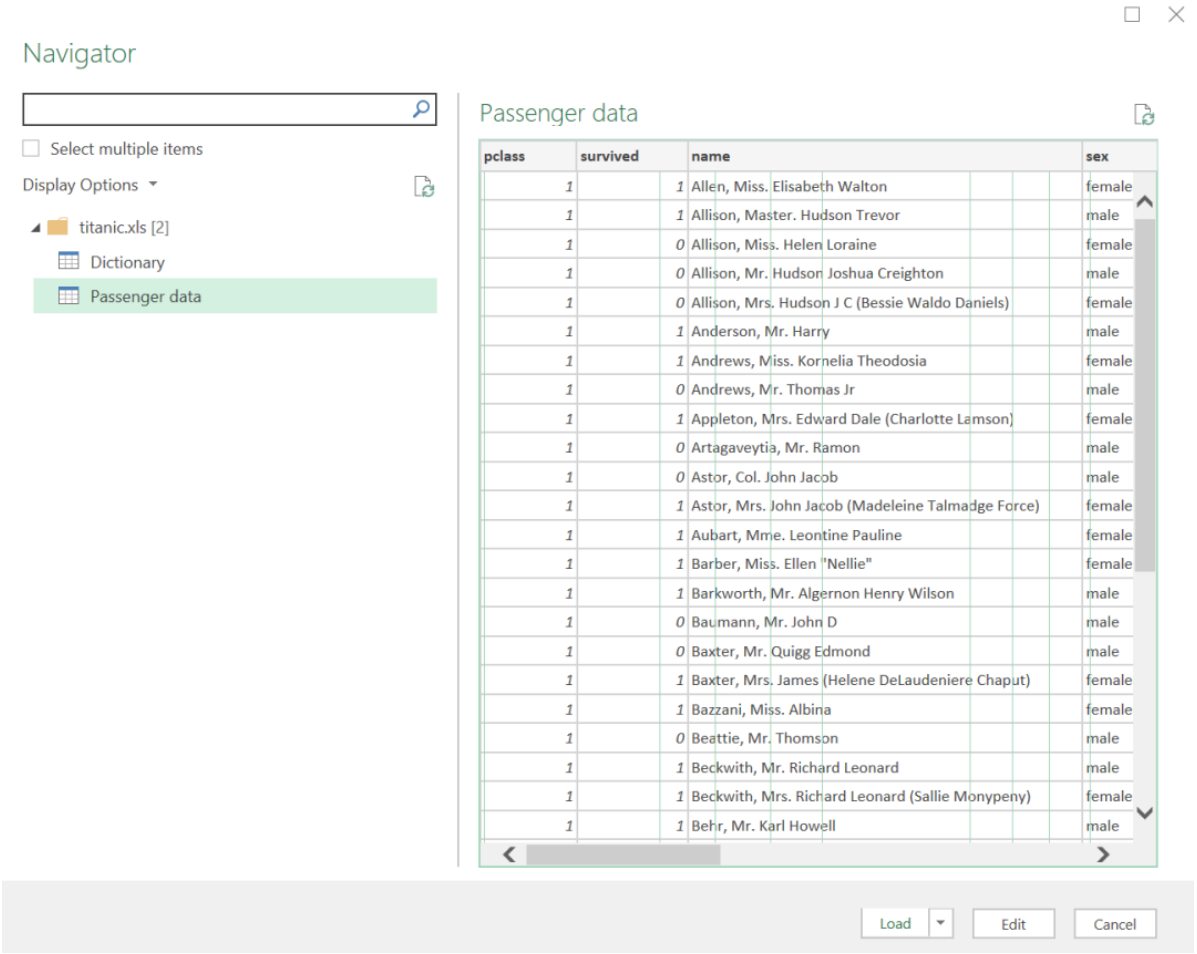
Load | Edit | Cancel

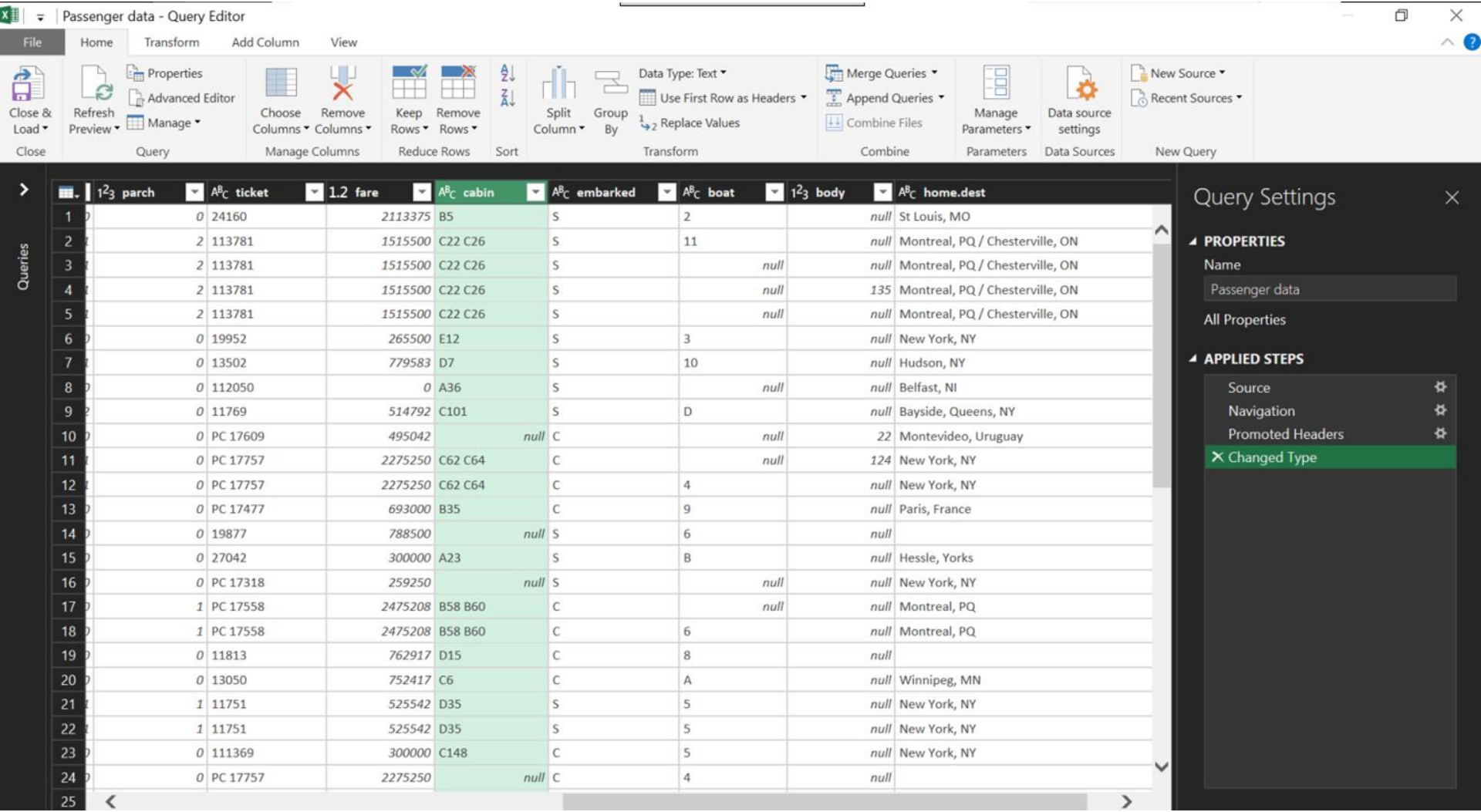
Importing data from another Excel workbook

Let's follow some simple steps to load and transform the data. While in a new workbook, follow these steps:

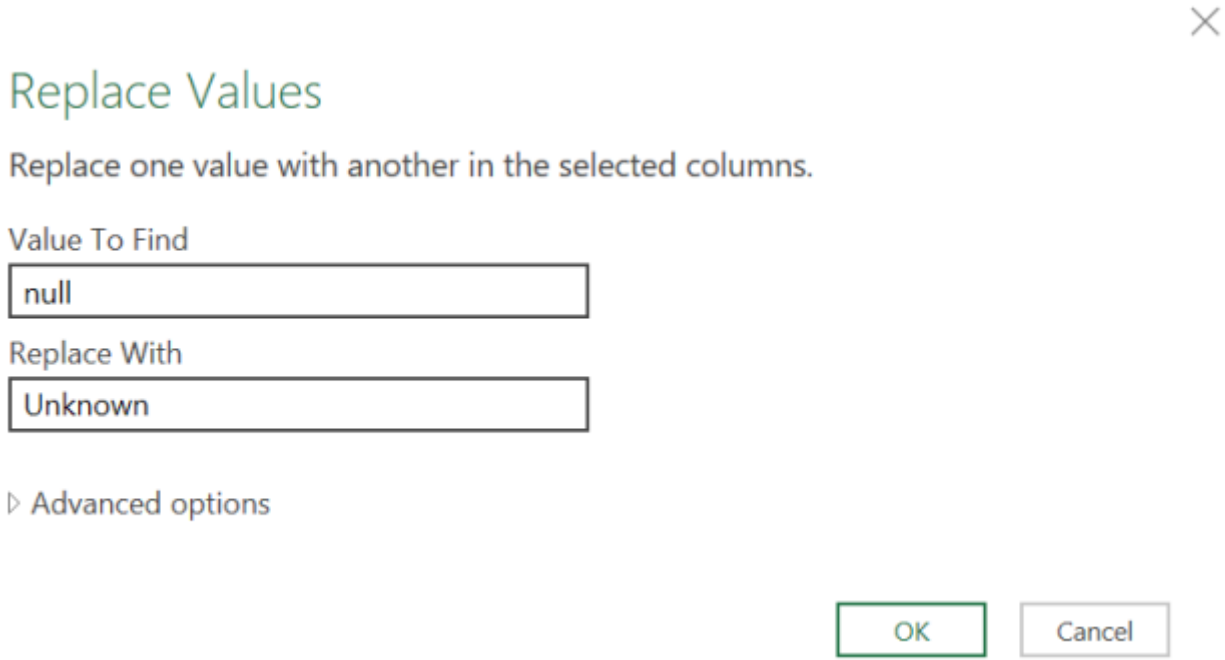
- Click on Data.
- Navigate to Get Data | From File | From Workbook, as shown in the following screenshot on next slide.

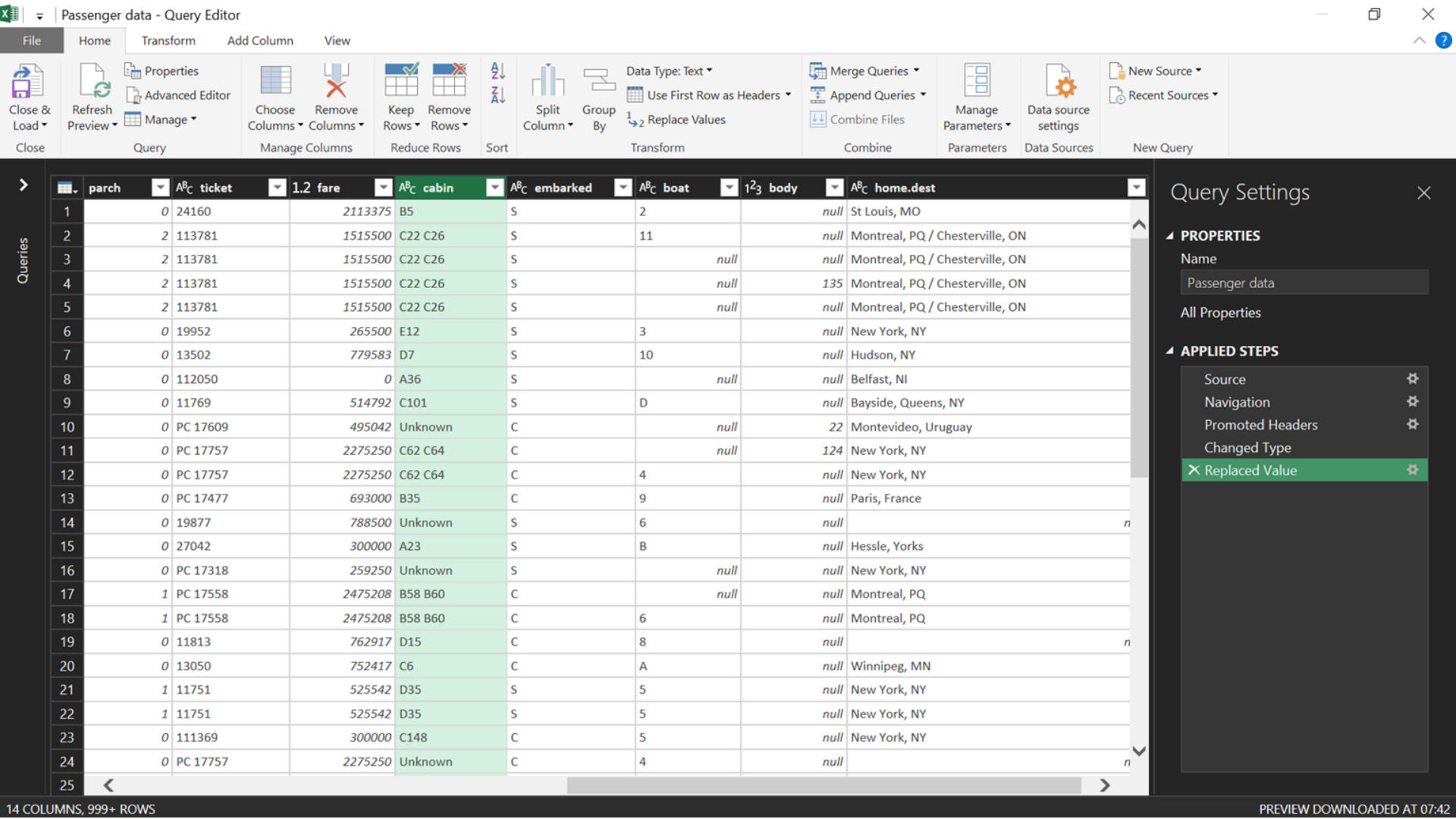

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Data Tools
What-If Analysis
Forecast Sheet
Forecast
Group
Ungroup
Subtotal
Outline
Data Analysis
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
From Workbook
From Text/CSV
From XML
From JSON
From Folder
From SharePoint Folder
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options

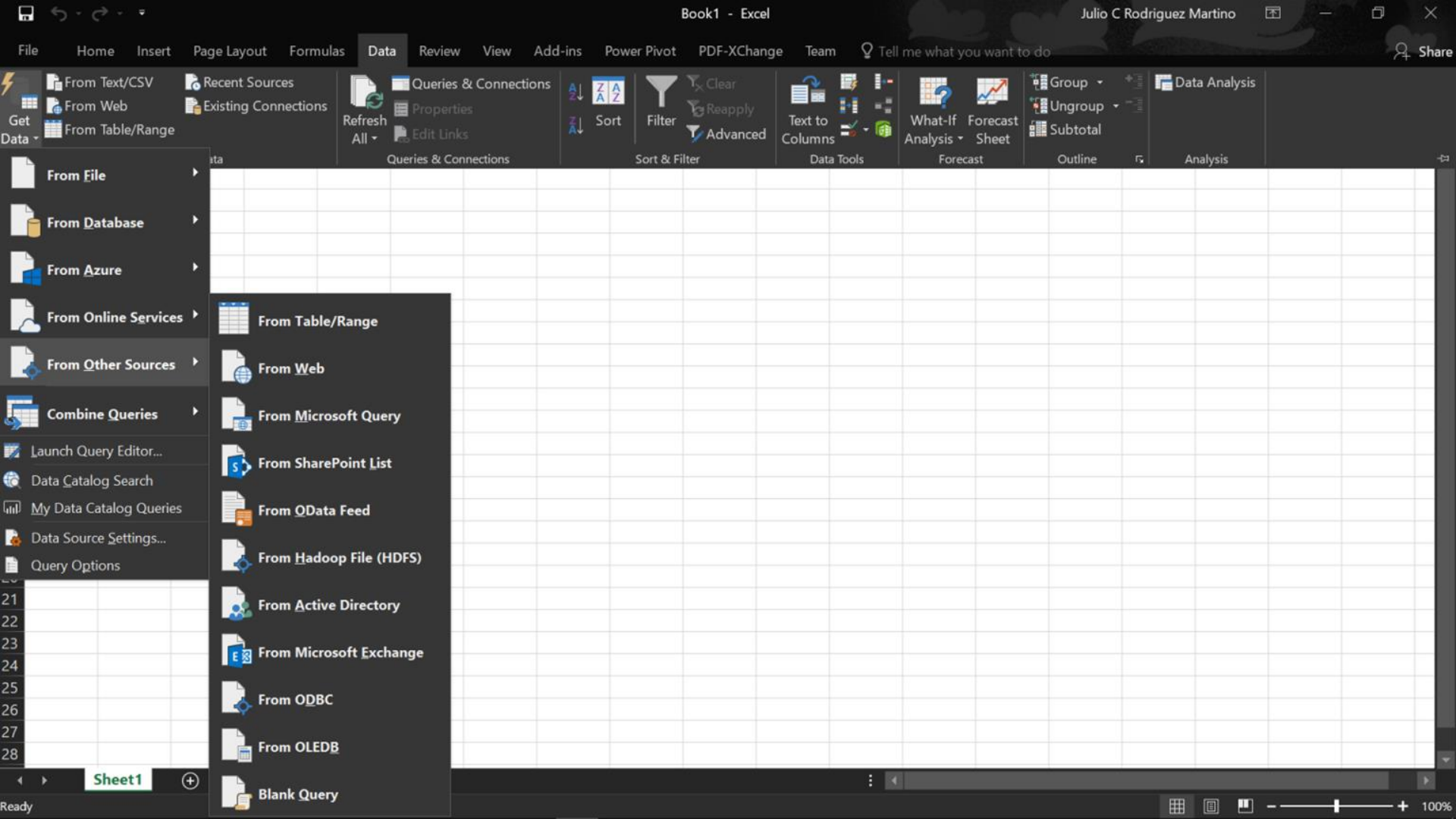

Navigator
Select multiple items
Display Options
titanic.xls [2]
Dictionary
Passenger data
Passenger data
pclass survived name sex
1 1 Allen, Miss. Elisabeth Walton female
1 1 Allison, Master. Hudson Trevor male
1 0 Allison, Miss. Helen Loraine female
1 0 Allison, Mr. Hudson Joshua Creighton male
1 0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
1 1 Anderson, Mr. Harry male
1 1 Andrews, Miss. Kornelia Theodosia female
1 0 Andrews, Mr. Thomas Jr male
1 1 Appleton, Mrs. Edward Dale (Charlotte Lamson) female
1 0 Artagaveytia, Mr. Ramon male
1 0 Astor, Col. John Jacob male
1 1 Astor, Mrs. John Jacob (Madeleine Talmadge Force) female
1 1 Aubart, Mme. Leontine Pauline female
1 1 Barber, Miss. Ellen "Nellie" female
1 1 Barkworth, Mr. Algernon Henry Wilson male
1 0 Baumann, Mr. John D male
1 0 Baxter, Mr. Quigg Edmond male
1 1 Baxter, Mrs. James (Helene DeLaudeniere Chaput) female
1 1 Bazzani, Miss. Albina female
1 0 Beattie, Mr. Thomson male
1 1 Beckwith, Mr. Richard Leonard male
1 1 Beckwith, Mrs. Richard Leonard (Sallie Monypeny) female
1 1 Behr, Mr. Karl Howell male
Load
Edit
Cancel


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Transform
Merge Queries
Append Queries
Combine Files
Combine
Manage Parameters
Parameters
Data source settings
Data Sources
New Source
Recent Sources
New Query
Queries
parch
ticket
fare
cabin
embarked
boat
body
home.dest
1 0 24160 2113375 B5 S 2 null St Louis, MO
2 2 113781 1515500 C22 C26 S 11 null Montreal, PQ / Chesterville, ON
3 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
4 2 113781 1515500 C22 C26 S null 135 Montreal, PQ / Chesterville, ON
5 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
6 0 19952 265500 E12 S 3 null New York, NY
7 0 13502 779583 D7 S 10 null Hudson, NY
8 0 112050 0 A36 S null null Belfast, NI
9 0 11769 514792 C101 S D null Bayside, Queens, NY
10 0 PC 17609 495042 null C null 22 Montevideo, Uruguay
11 0 PC 17757 2275250 C62 C64 C null 124 New York, NY
12 0 PC 17757 2275250 C62 C64 C 4 null New York, NY
13 0 PC 17477 693000 B35 C 9 null Paris, France
14 0 19877 788500 null S 6 null
15 0 27042 300000 A23 S B null Hessle, Yorks
16 0 PC 17318 259250 null S null null New York, NY
17 1 PC 17558 2475208 B58 B60 C null null Montreal, PQ
18 1 PC 17558 2475208 B58 B60 C 6 null Montreal, PQ
19 0 11813 762917 D15 C 8 null
20 0 13050 752417 C6 C A null Winnipeg, MN
21 1 11751 525542 D35 S 5 null New York, NY
22 1 11751 525542 D35 S 5 null New York, NY
23 0 111369 300000 C148 C 5 null New York, NY
24 0 PC 17757 2275250 null C 4 null
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type

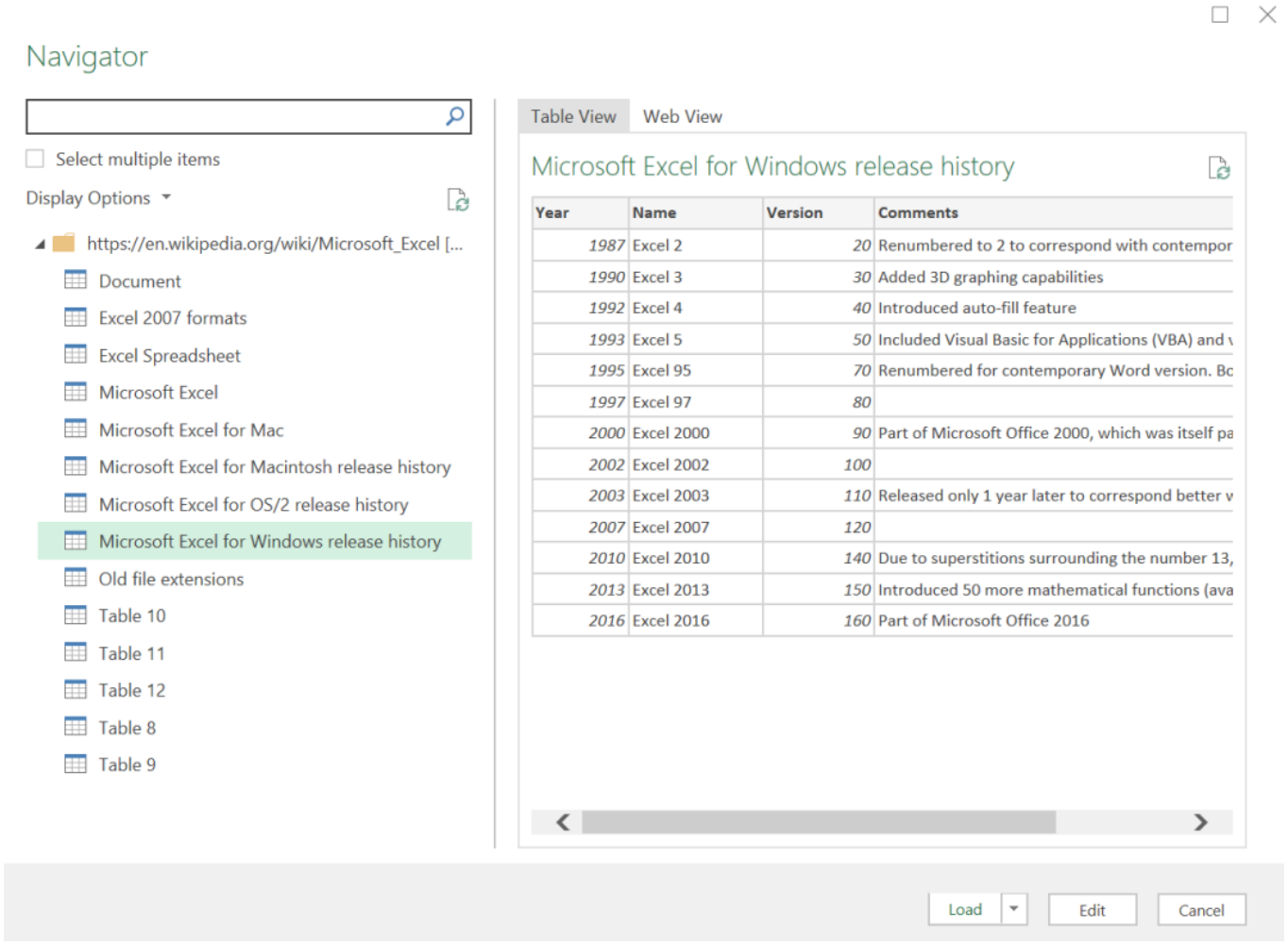
- Click on Replace Values in the Transform menu.
- You will get a pop-up dialog where you can tell Excel to replace null with Unknown, as shown in the following screenshot:

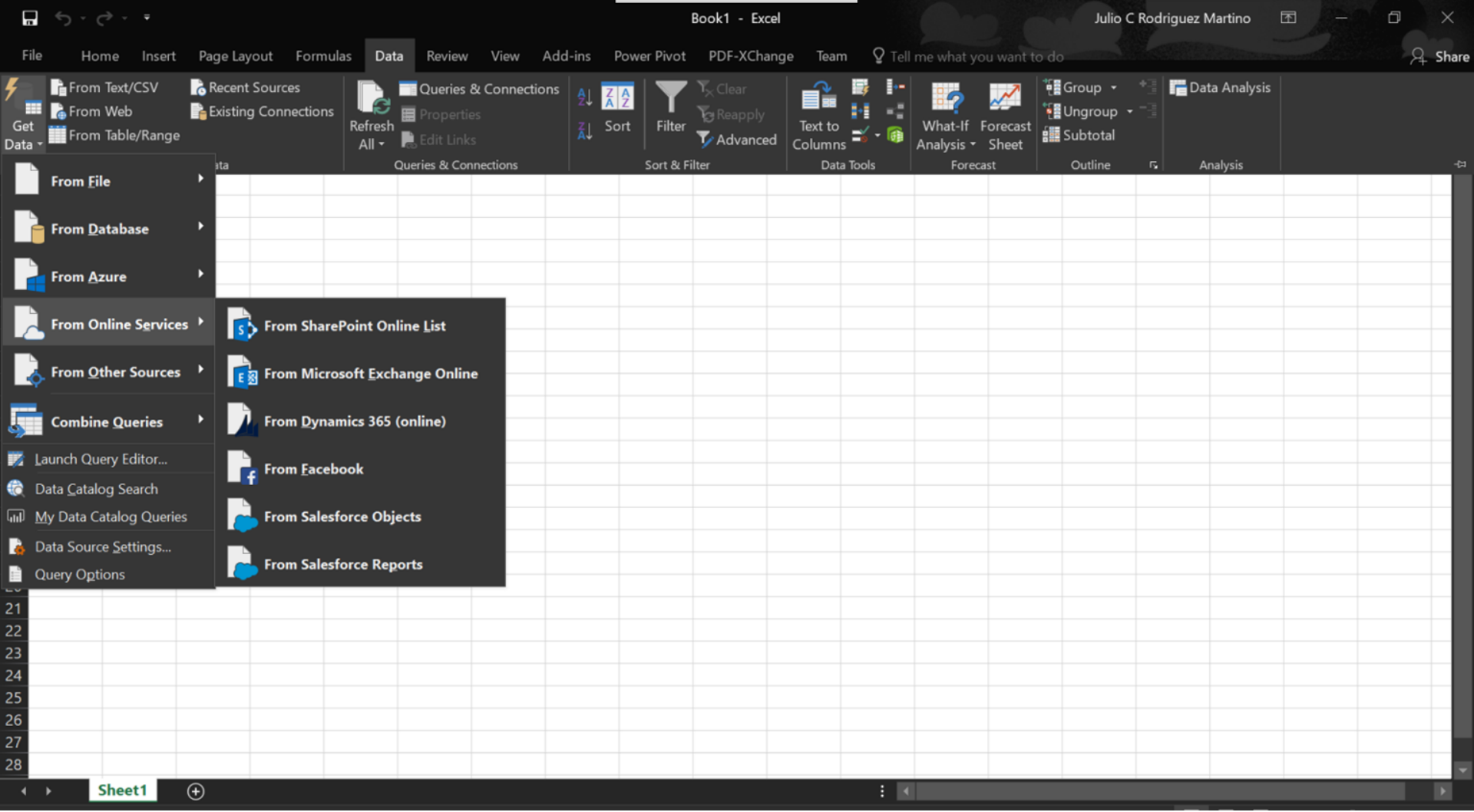



Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Transform
Merge Queries
Append Queries
Combine Files
Combine
Manage Parameters
Parameters
Data source settings
Data Sources
New Source
Recent Sources
New Query
Queries
parch
ticket
fare
cabin
embarked
boat
body
home.dest
1 0 24160 2113375 B5 S 2 null St Louis, MO
2 2 113781 1515500 C22 C26 S 11 null Montreal, PQ / Chesterville, ON
3 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
4 2 113781 1515500 C22 C26 S null 135 Montreal, PQ / Chesterville, ON
5 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
6 0 19952 265500 E12 S 3 null New York, NY
7 0 13502 779583 D7 S 10 null Hudson, NY
8 0 112050 0 A36 S null null Belfast, NI
9 0 11769 514792 C101 S D null Bayside, Queens, NY
10 0 PC 17609 495042 Unknown C null 22 Montevideo, Uruguay
11 0 PC 17757 2275250 C62 C64 C null 124 New York, NY
12 0 PC 17757 2275250 C62 C64 C 4 null New York, NY
13 0 PC 17477 693000 B35 C 9 null Paris, France
14 0 19877 788500 Unknown S 6 null n
15 0 27042 300000 A23 S B null Hessle, Yorks
16 0 PC 17318 259250 Unknown S null null New York, NY
17 1 PC 17558 2475208 B58 B60 C null null Montreal, PQ
18 1 PC 17558 2475208 B58 B60 C 6 null Montreal, PQ
19 0 11813 762917 D15 C 8 null n
20 0 13050 752417 C6 C A null Winnipeg, MN
21 1 11751 525542 D35 S 5 null New York, NY
22 1 11751 525542 D35 S 5 null New York, NY
23 0 111369 300000 C148 C 5 null New York, NY
24 0 PC 17757 2275250 Unknown C 4 null n
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Replaced Value
14 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 07:42

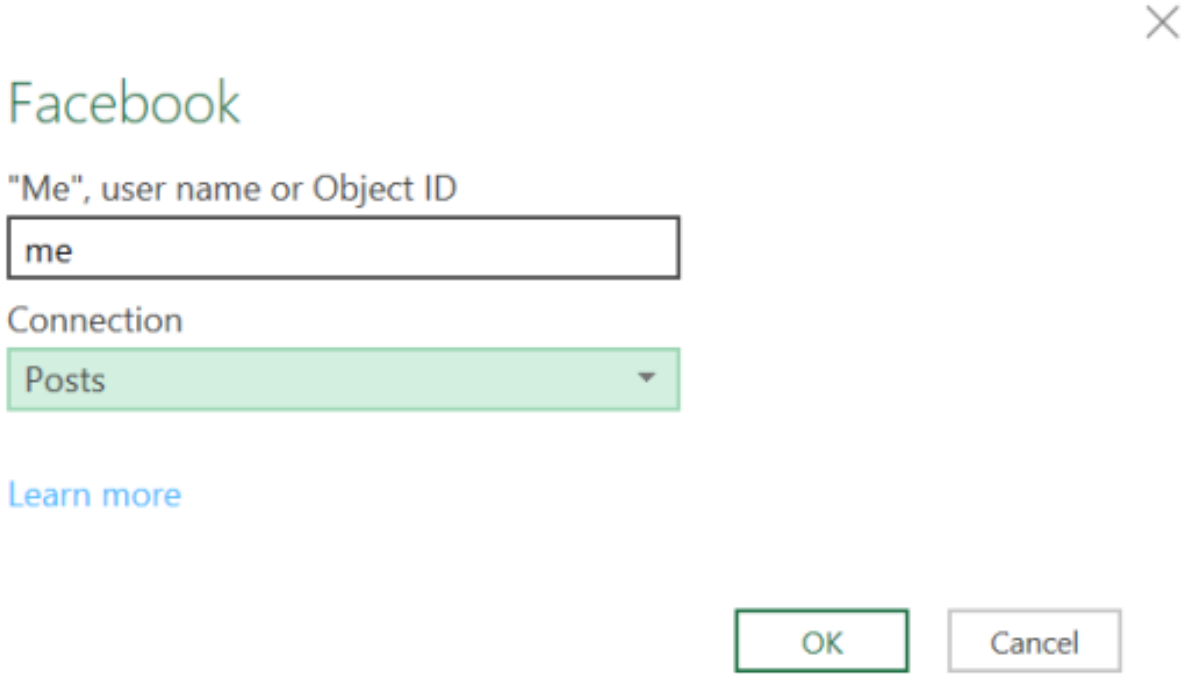

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Queries & Connections
Sort
Filter
Clear
Reapply
Advanced
Sort & Filter
Text to Columns
Data Tools
What-If Analysis
Forecast Sheet
Forecast
Group
Ungroup
Subtotal
Outline
Data Analysis
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options
From Table/Range
From Web
From Microsoft Query
From SharePoint List
From OData Feed
From Hadoop File (HDFS)
From Active Directory
From Microsoft Exchange
From ODBC
From OLEDB
Blank Query
Sheet1
Ready
100%

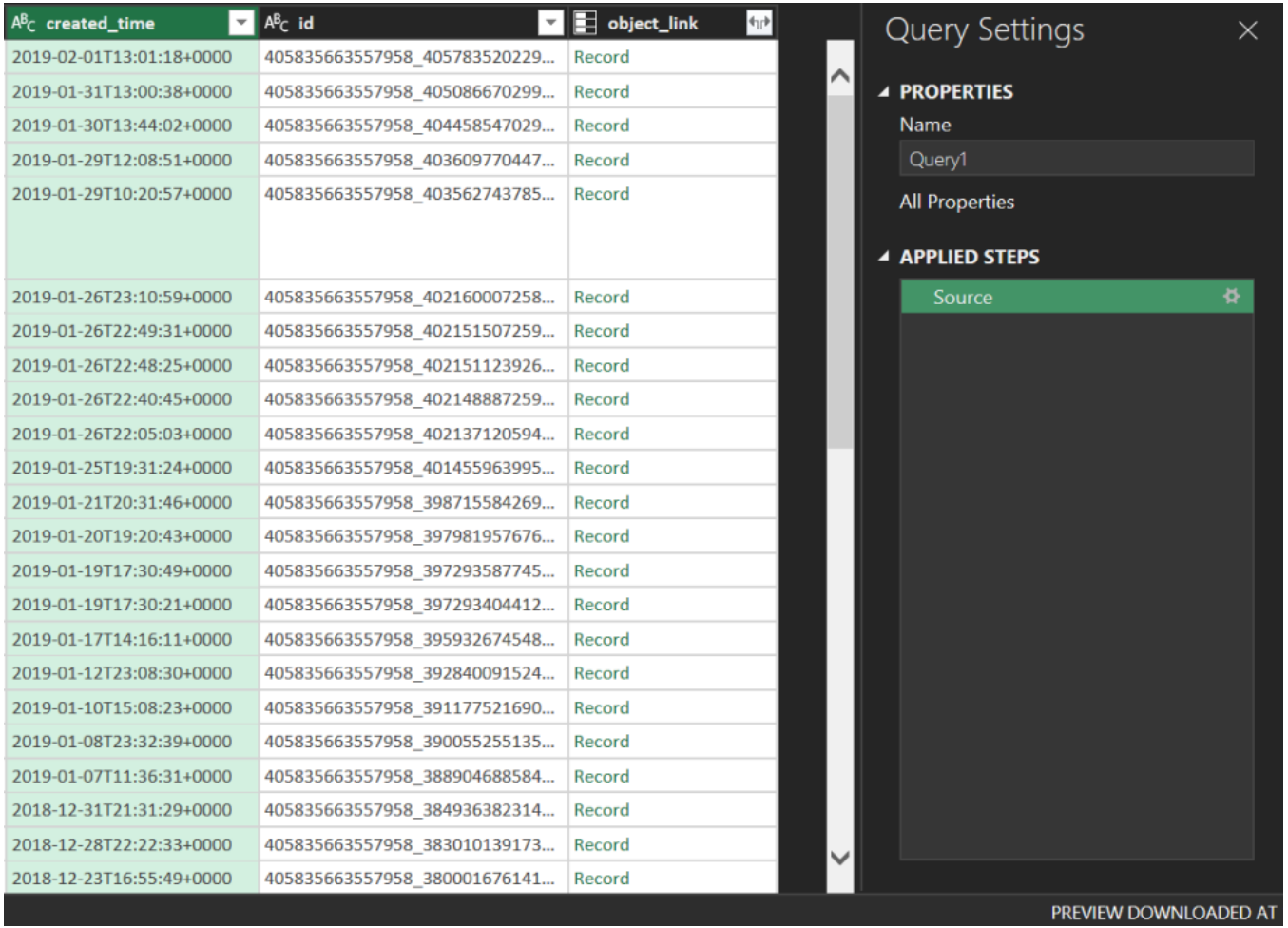
Importing data from a web page


Navigator
Select multiple items
Display Options
https://en.wikipedia.org/wiki/Microsoft_Excel [...
Document
Excel 2007 formats
Excel Spreadsheet
Microsoft Excel
Microsoft Excel for Mac
Microsoft Excel for Macintosh release history
Microsoft Excel for OS/2 release history
Microsoft Excel for Windows release history
Old file extensions
Table 10
Table 11
Table 12
Table 8
Table 9
Table View
Web View
Microsoft Excel for Windows release history
Year
Name
Version
Comments
1987 Excel 2 20 Renumbered to 2 to correspond with contempor
1990 Excel 3 30 Added 3D graphing capabilities
1992 Excel 4 40 Introduced auto-fill feature
1993 Excel 5 50 Included Visual Basic for Applications (VBA) and
1995 Excel 95 70 Renumbered for contemporary Word version. Bc
1997 Excel 97 80
2000 Excel 2000 90 Part of Microsoft Office 2000, which was itself pa
2002 Excel 2002 100
2003 Excel 2003 110 Released only 1 year later to correspond better
2007 Excel 2007 120
2010 Excel 2010 140 Due to superstitions surrounding the number 13,
2013 Excel 2013 150 Introduced 50 more mathematical functions (ava
2016 Excel 2016 160 Part of Microsoft Office 2016
Load
Edit
Cancel

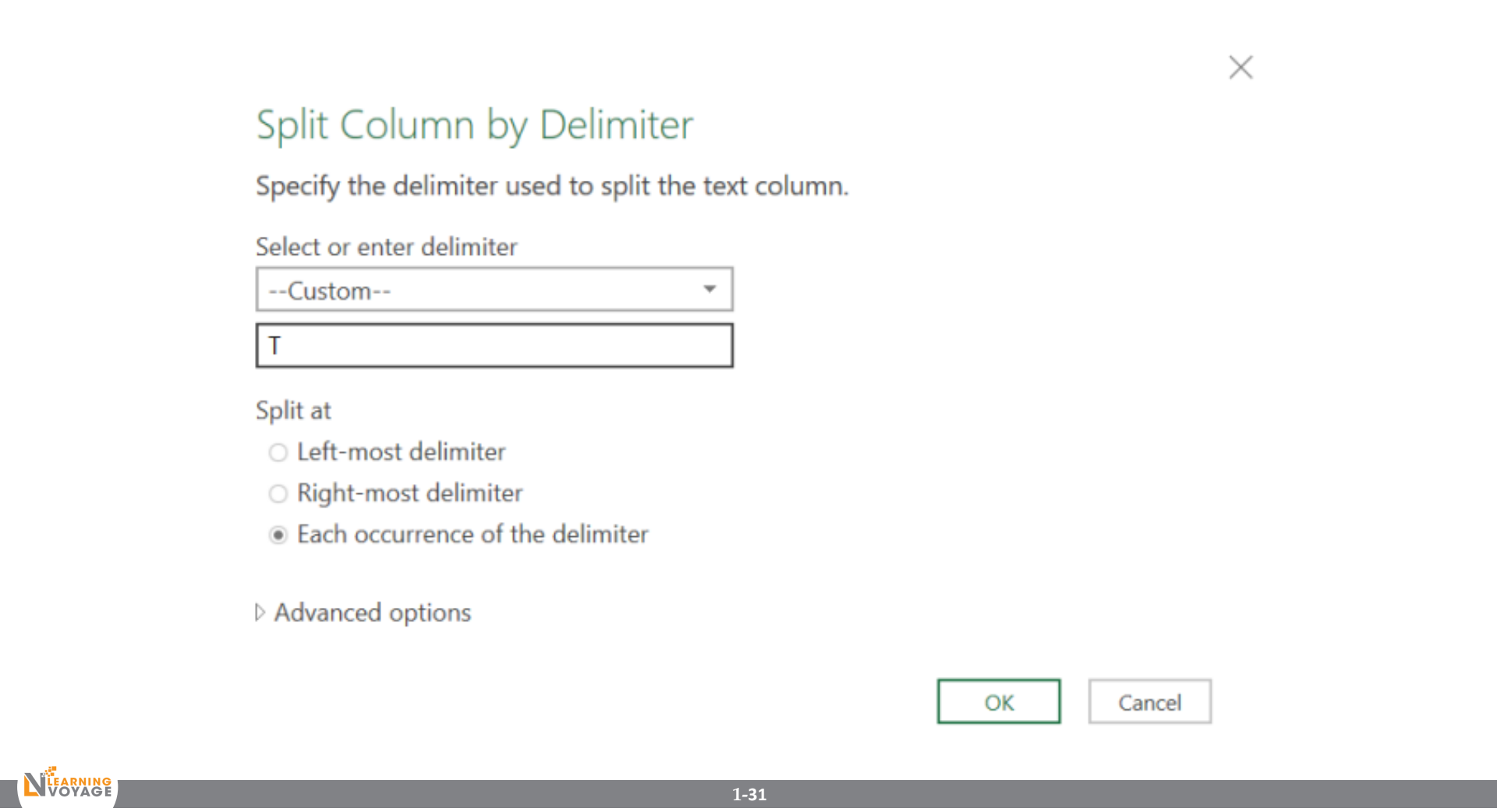

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Group
Ungroup
Subtotal
Data Analysis
Queries & Connections
Sort & Filter
Data Tools
Forecast
Outline
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options
From SharePoint Online List
From Microsoft Exchange Online
From Dynamics 365 (online)
From Facebook
From Salesforce Objects
From Salesforce Reports
Sheet1

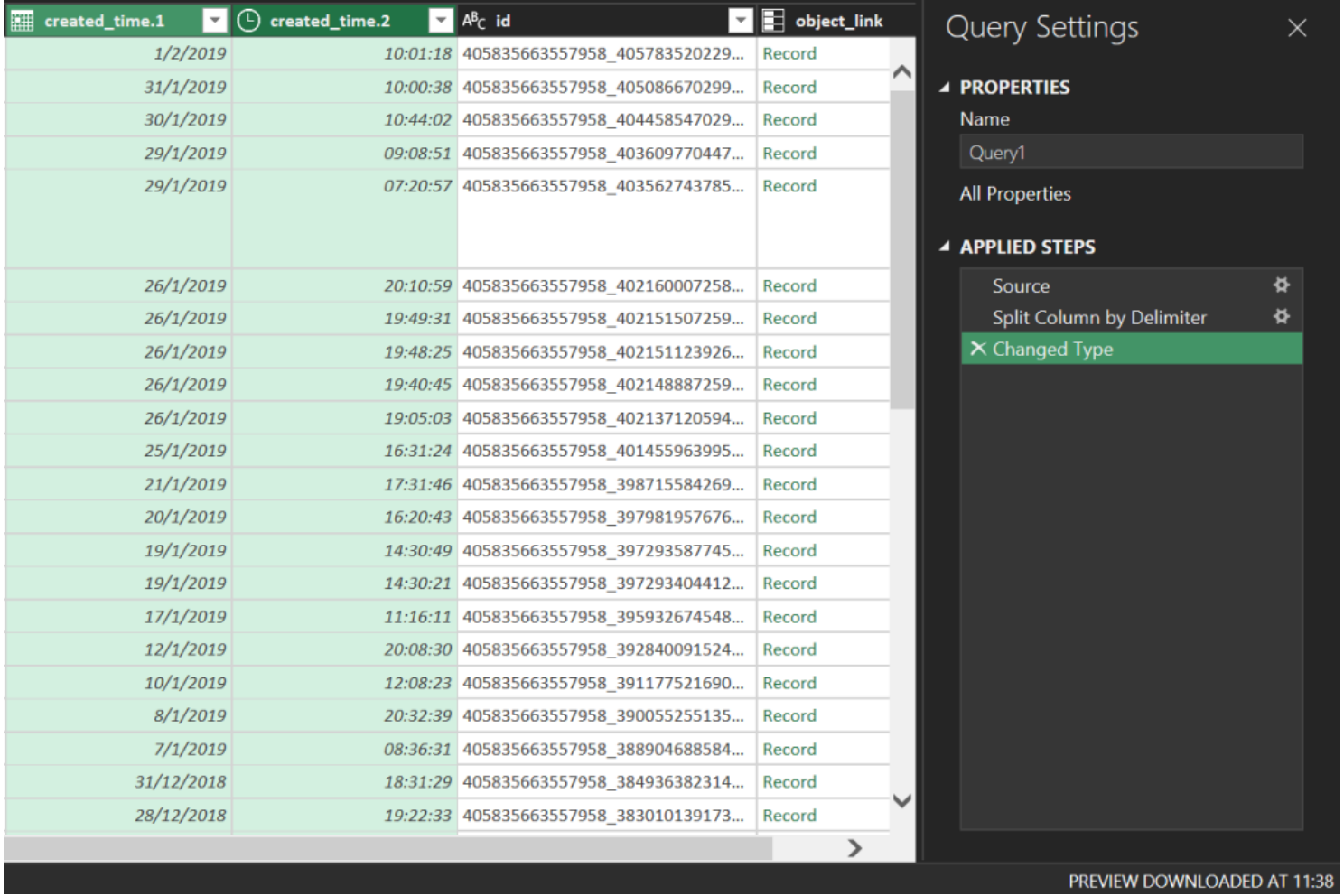
Importing data from Facebook





created_time	id	object_link
2019-02-01T13:01:18+0000	405835663557958_405783520229...	Record
2019-01-31T13:00:38+0000	405835663557958_405086670299...	Record
2019-01-30T13:44:02+0000	405835663557958_404458547029...	Record
2019-01-29T12:08:51+0000	405835663557958_403609770447...	Record
2019-01-29T10:20:57+0000	405835663557958_403562743785...	Record
2019-01-26T23:10:59+0000	405835663557958_402160007258...	Record
2019-01-26T22:49:31+0000	405835663557958_402151507259...	Record
2019-01-26T22:48:25+0000	405835663557958_402151123926...	Record
2019-01-26T22:40:45+0000	405835663557958_402148887259...	Record
2019-01-26T22:05:03+0000	405835663557958_402137120594...	Record
2019-01-25T19:31:24+0000	405835663557958_401455963995...	Record
2019-01-21T20:31:46+0000	405835663557958_398715584269...	Record
2019-01-20T19:20:43+0000	405835663557958_397981957676...	Record
2019-01-19T17:30:49+0000	405835663557958_397293587745...	Record
2019-01-19T17:30:21+0000	405835663557958_397293404412...	Record
2019-01-17T14:16:11+0000	405835663557958_395932674548...	Record
2019-01-12T23:08:30+0000	405835663557958_392840091524...	Record
2019-01-10T15:08:23+0000	405835663557958_391177521690...	Record
2019-01-08T23:32:39+0000	405835663557958_390055255135...	Record
2019-01-07T11:36:31+0000	405835663557958_388904688584...	Record
2018-12-31T21:31:29+0000	405835663557958_384936382314...	Record
2018-12-28T22:22:33+0000	405835663557958_383010139173...	Record
2018-12-23T16:55:49+0000	405835663557958_380001676141...	Record


Split Column by Delimiter
Specify the delimiter used to split the text column.
Select or enter delimiter
--Custom--
T
Split at
Left-most delimiter
Right-most delimiter
Each occurrence of the delimiter
Advanced options
OK
Cancel


created_time.1
created_time.2
id
object_link
1/2/2019 10:01:18 405835663557958_405783520229... Record
31/1/2019 10:00:38 405835663557958_405086670299... Record
30/1/2019 10:44:02 405835663557958_404458547029... Record
29/1/2019 09:08:51 405835663557958_403609770447... Record
29/1/2019 07:20:57 405835663557958_403562743785... Record
26/1/2019 20:10:59 405835663557958_402160007258... Record
26/1/2019 19:49:31 405835663557958_402151507259... Record
26/1/2019 19:48:25 405835663557958_402151123926... Record
26/1/2019 19:40:45 405835663557958_402148887259... Record
26/1/2019 19:05:03 405835663557958_402137120594... Record
25/1/2019 16:31:24 405835663557958_401455963995... Record
21/1/2019 17:31:46 405835663557958_398715584269... Record
20/1/2019 16:20:43 405835663557958_397981957676... Record
19/1/2019 14:30:49 405835663557958_397293587745... Record
19/1/2019 14:30:21 405835663557958_397293404412... Record
17/1/2019 11:16:11 405835663557958_395932674548... Record
12/1/2019 20:08:30 405835663557958_392840091524... Record
10/1/2019 12:08:23 405835663557958_391177521690... Record
8/1/2019 20:32:39 405835663557958_390055255135... Record
7/1/2019 08:36:31 405835663557958_388904688584... Record
31/12/2018 18:31:29 405835663557958_384936382314... Record
28/12/2018 19:22:33 405835663557958_383010139173... Record
Query Settings
PROPERTIES
Name
Query1
All Properties
APPLIED STEPS
Source
Split Column by Delimiter
Changed Type
PREVIEW DOWNLOADED AT 11:38

Importing data from a JSON file

- JSON is a standard format for sharing data, since it uses text fields that can be read by a human being.
- It is used by most web applications for data input and output.
- In our example, we will use the Azure Text Analytics API.
- Given a sentence, this service can identify the text sentiment and the language and extract keywords, among other things.

I had a wonderful trip to Seattle and enjoyed seeing the Space Needle!

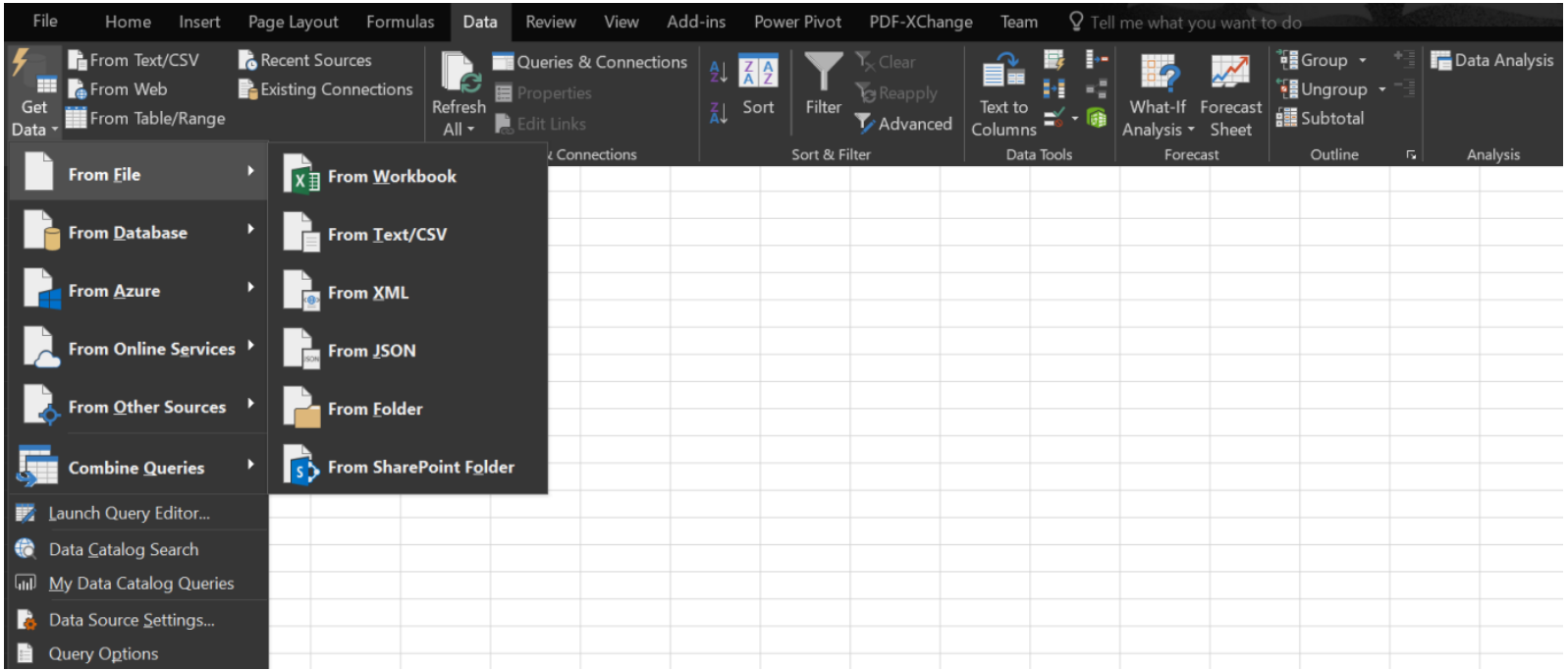
Analyze

Example - English - Positive | Example - English - Negative | Example - Spanish - Positive | Example - Spanish - Negative

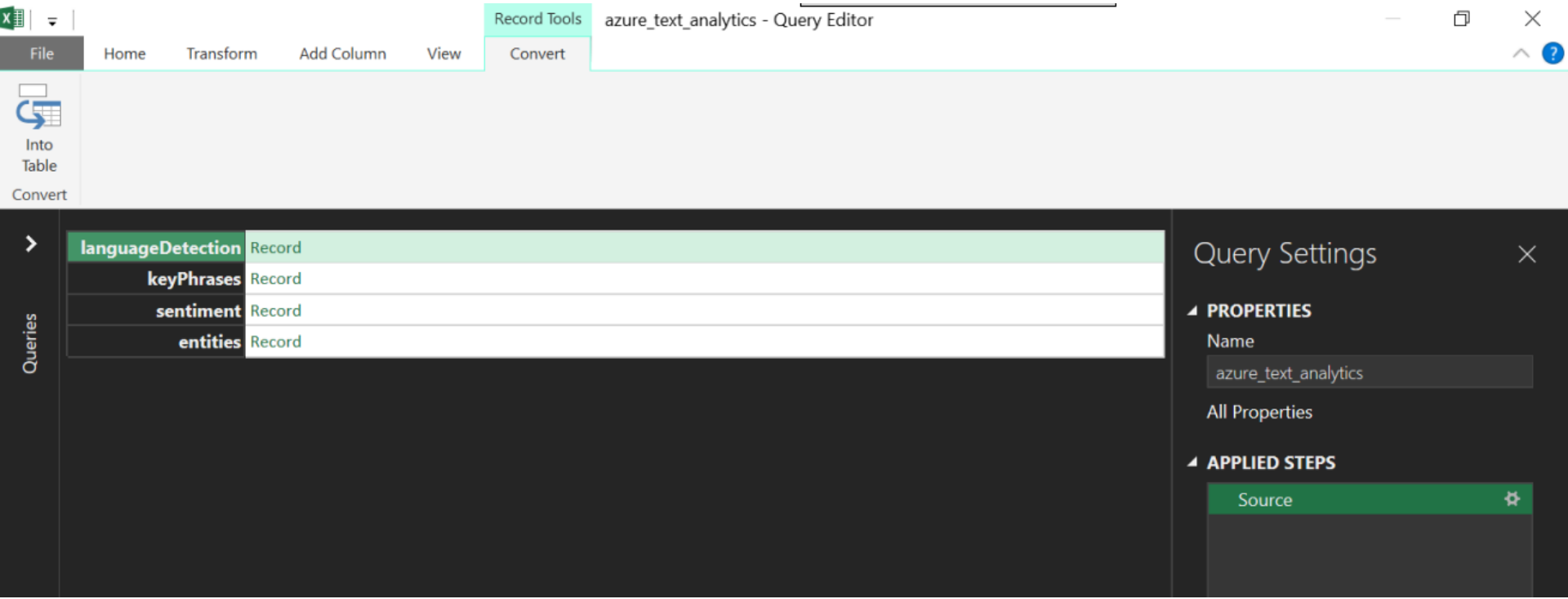
Analyzed text | JSON


```
{
  "languageDetection": {
    "documents": [
      {
        "id": "cb64821e-57fd-47e9-b745-0ad7d15b3ece",
        "detectedLanguages": [
          {
            "name": "English",
            "iso6391Name": "en",
            "score": 1.0
          }
        ]
      }
    ],
    "errors": []
  },
  "keyPhrases": {
    "documents": [
      {
        "id": "cb64821e-57fd-47e9-b745-0ad7d15b3ece",
        "keyPhrases": [
          "Seattle",
          "wonderful trip",
```

- Click on Data.
- Navigate to Get Data | From File | From JSON, as shown in the following screenshot:

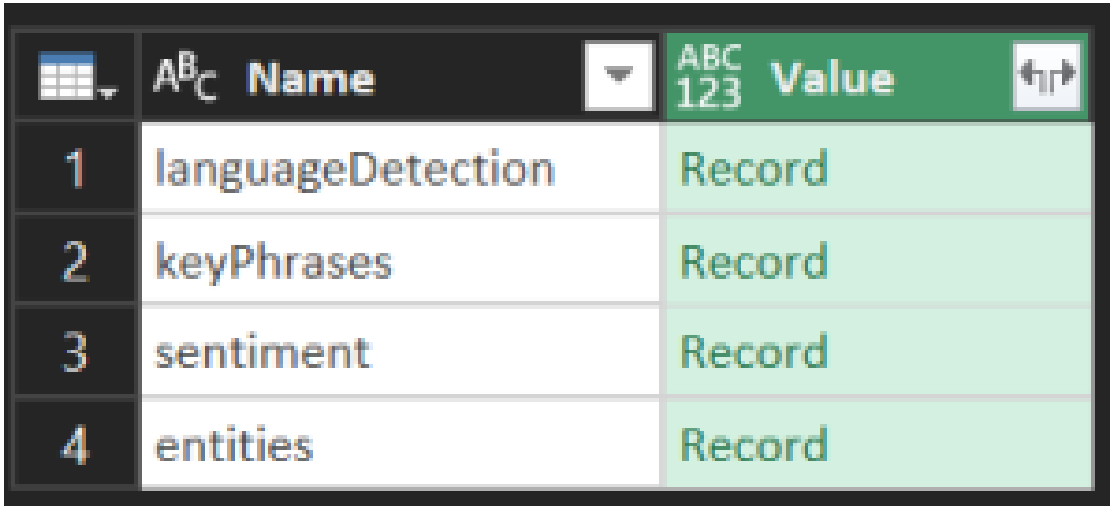


- You will get a preview showing the main fields in the JSON structure, as shown in the following screenshot:

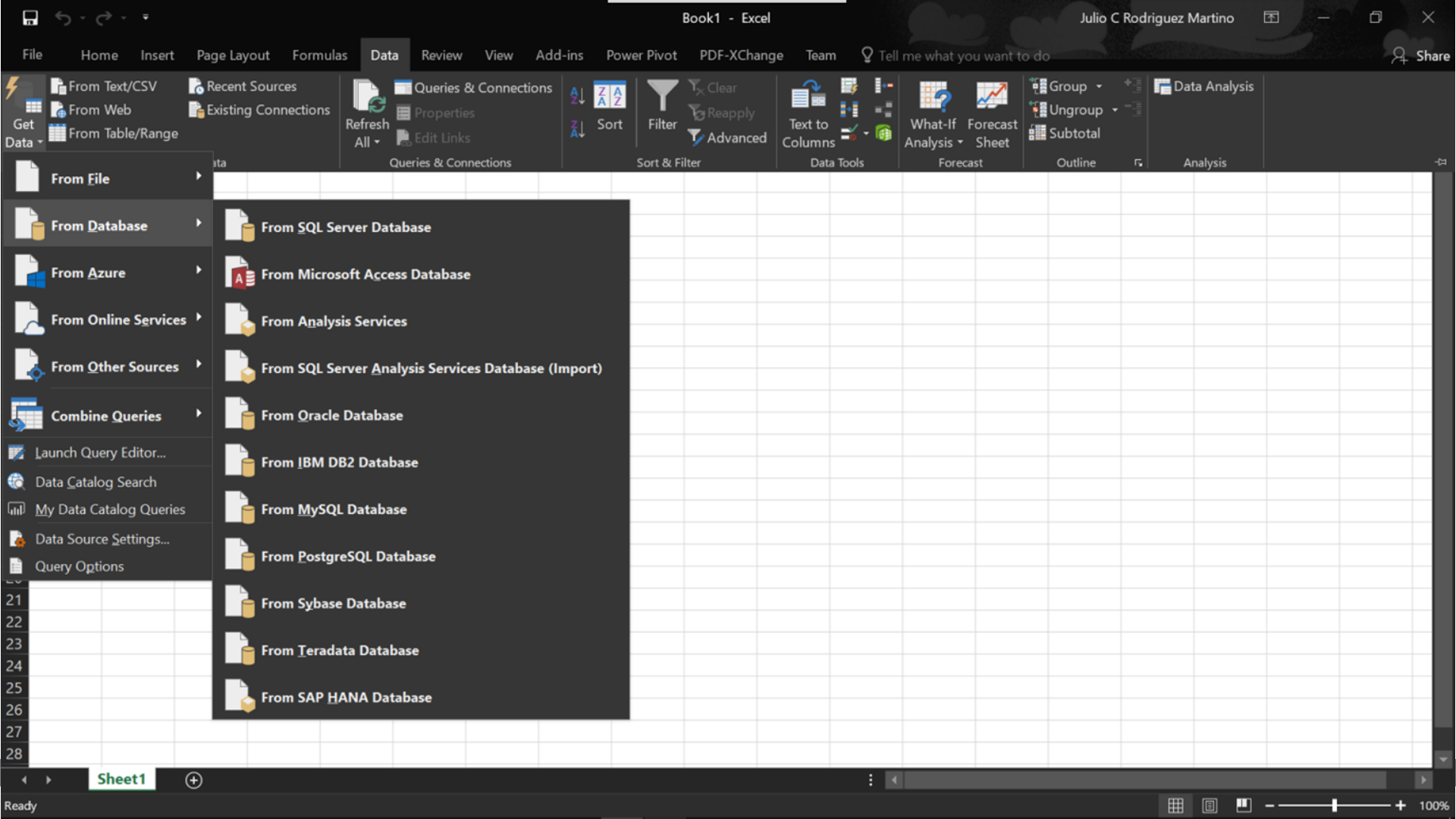


Importing data from a JSON file

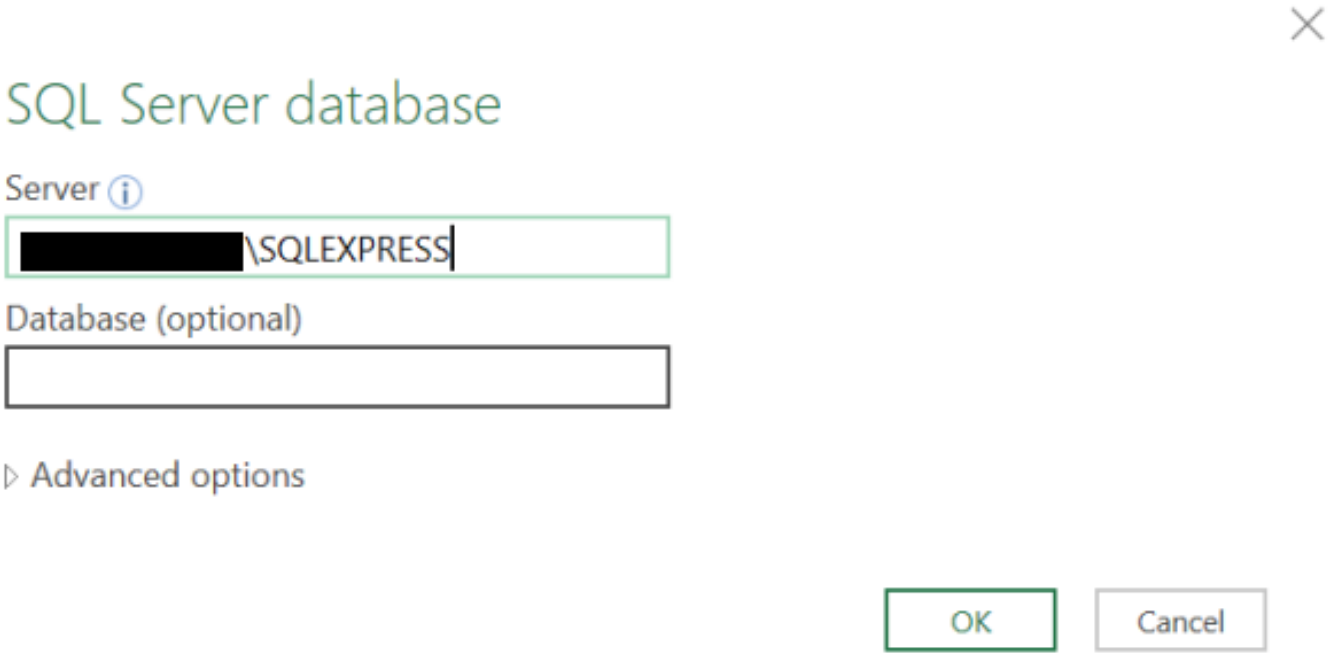
- Click on Into Table to convert the entries into regular Excel tables, as shown in the following screenshot



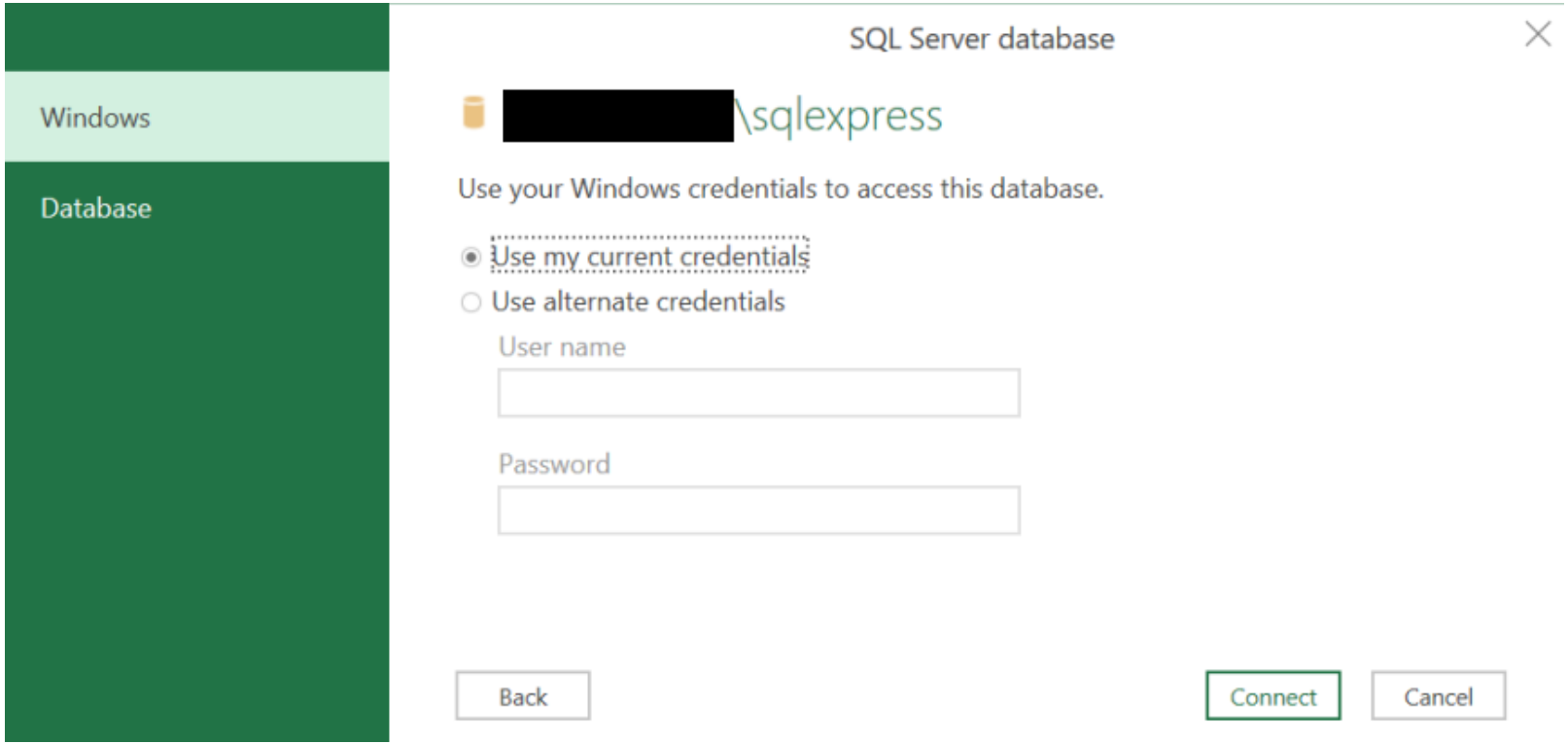
	Name	Value
1	languageDetection	Record
2	keyPhrases	Record
3	sentiment	Record
4	entities	Record

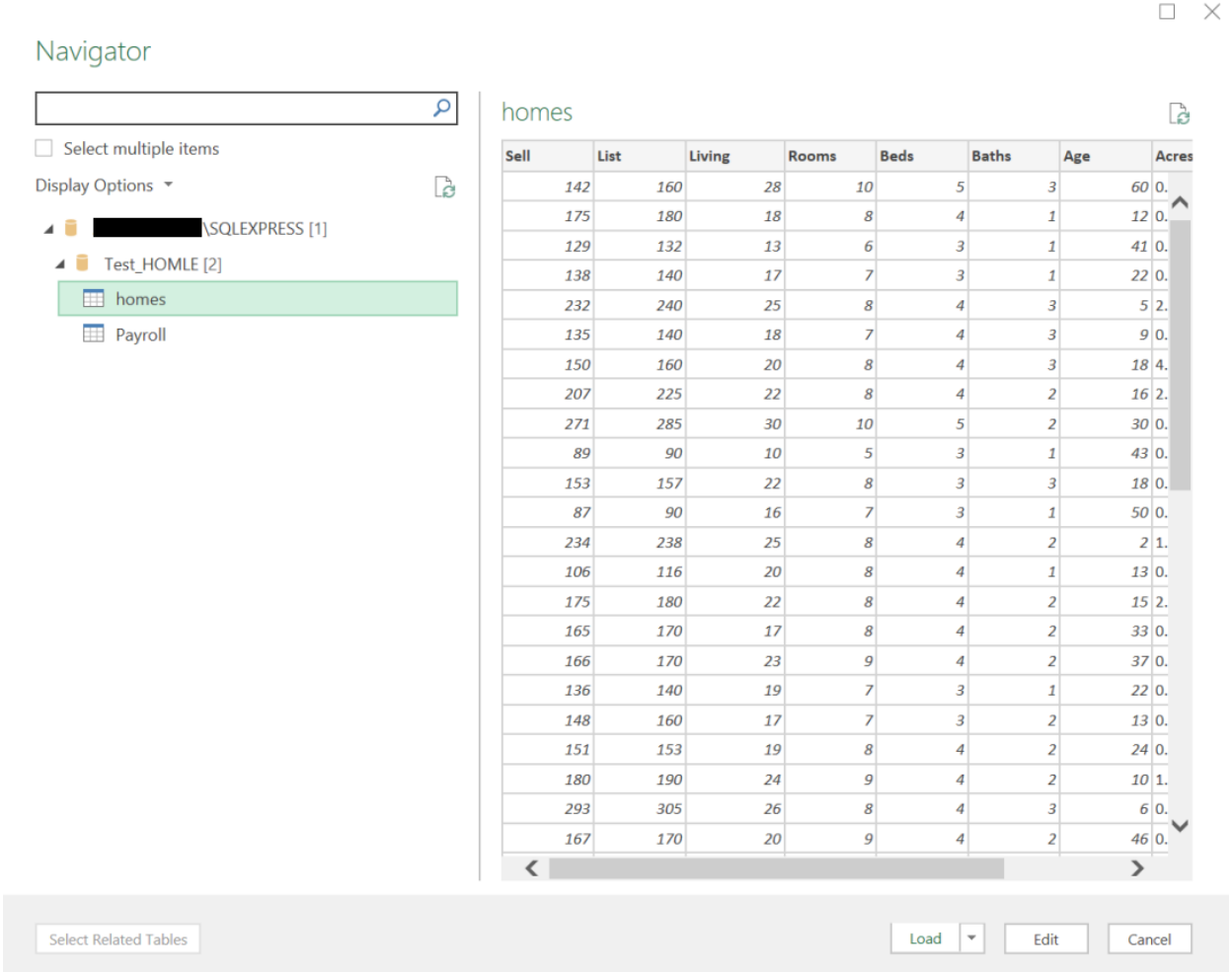

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Group
Ungroup
Subtotal
Data Analysis
Queries & Connections
Sort & Filter
Data Tools
Forecast
Outline
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options
From SQL Server Database
From Microsoft Access Database
From Analysis Services
From SQL Server Analysis Services Database (Import)
From Oracle Database
From IBM DB2 Database
From MySQL Database
From PostgreSQL Database
From Sybase Database
From Teradata Database
From SAP HANA Database
Sheet1
Ready
100%

Importing data from a database



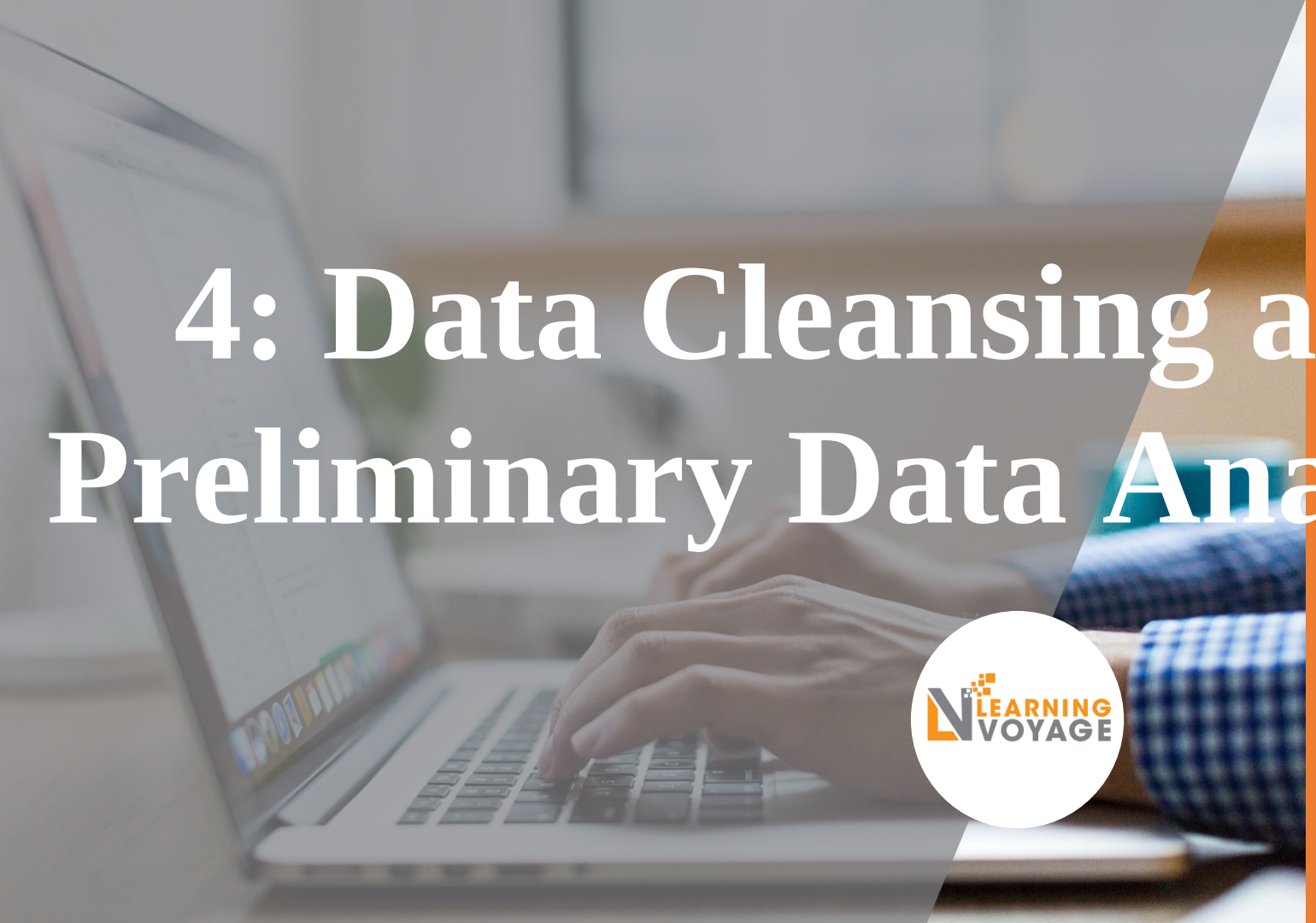
Importing data from a database




Navigator
Select multiple items
Display Options
[redacted]\SQLEXPRESS [1]
Test_HOMLE [2]
homes
Payroll
homes
Sell List Living Rooms Beds Baths Age Acres
142 160 28 10 5 3 60 0.
175 180 18 8 4 1 12 0.
129 132 13 6 3 1 41 0.
138 140 17 7 3 1 22 0.
232 240 25 8 4 3 5 2.
135 140 18 7 4 3 9 0.
150 160 20 8 4 3 18 4.
207 225 22 8 4 2 16 2.
271 285 30 10 5 2 30 0.
89 90 10 5 3 1 43 0.
153 157 22 8 3 3 18 0.
87 90 16 7 3 1 50 0.
234 238 25 8 4 2 2 1.
106 116 20 8 4 1 13 0.
175 180 22 8 4 2 15 2.
165 170 17 8 4 2 33 0.
166 170 23 9 4 2 37 0.
136 140 19 7 3 1 22 0.
148 160 17 7 3 2 13 0.
151 153 19 8 4 2 24 0.
180 190 24 9 4 2 10 1.
293 305 26 8 4 3 6 0.
167 170 20 9 4 2 46 0.
Select Related Tables
Load
Edit
Cancel

Summary

- In this lesson, we described different methods of inputting information into an Excel spreadsheet, going beyond what can be done by hand-typing data.
- A variety of file types, web data sources, and databases can be analyzed from within Excel by using Power Query and Query Editor to extract, transform, and load data.
- I encourage you to explore other data sources, since the loading procedure is very similar.



4: Data Cleansing and Preliminary Data Analysis

LEARNING VOYAGE

Data Cleansing and Preliminary Data Analysis

In this lesson, we will cover the following topics:

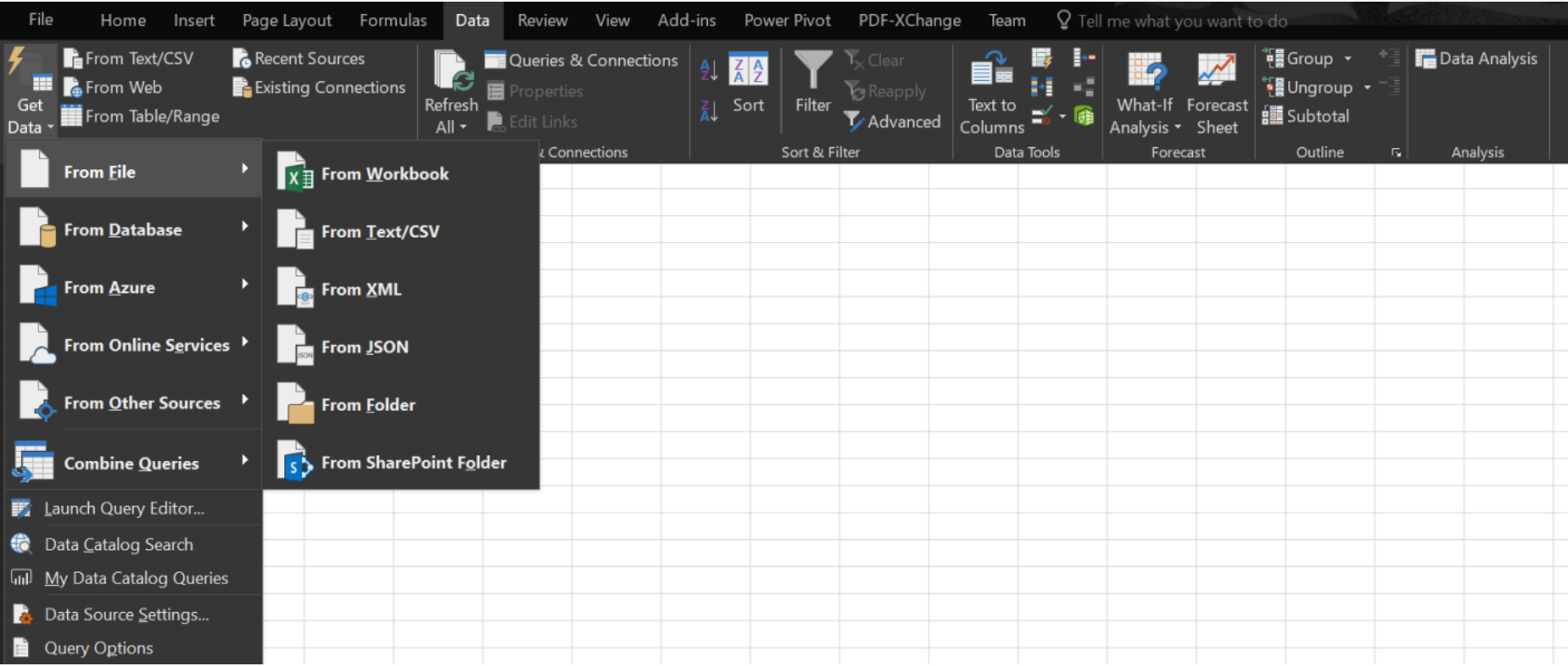
- Cleansing data
- Visualizing data for preliminary analysis
- Understanding unbalanced datasets

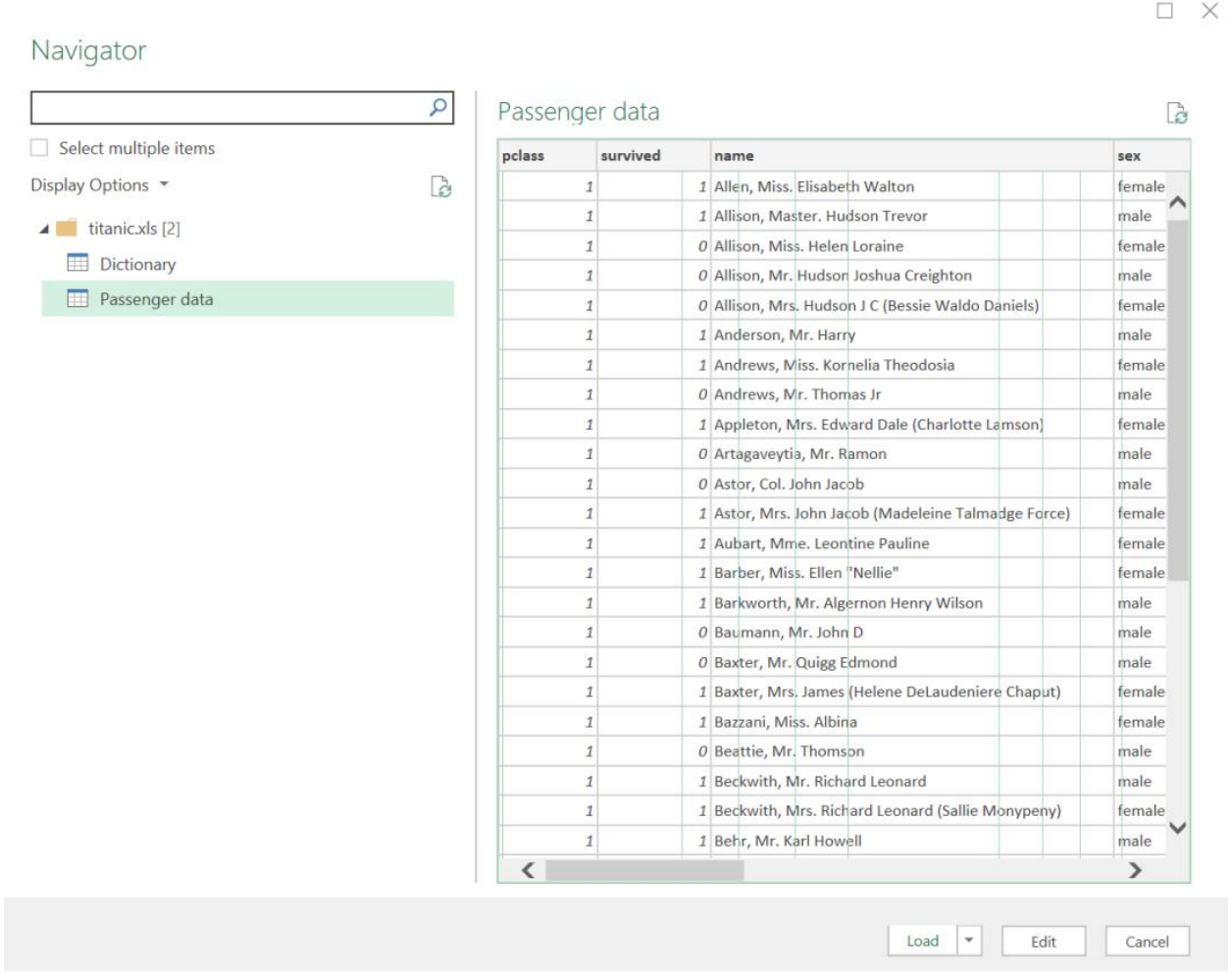
Technical requirements

- You will need to download the titanic.xls file

Cleansing data

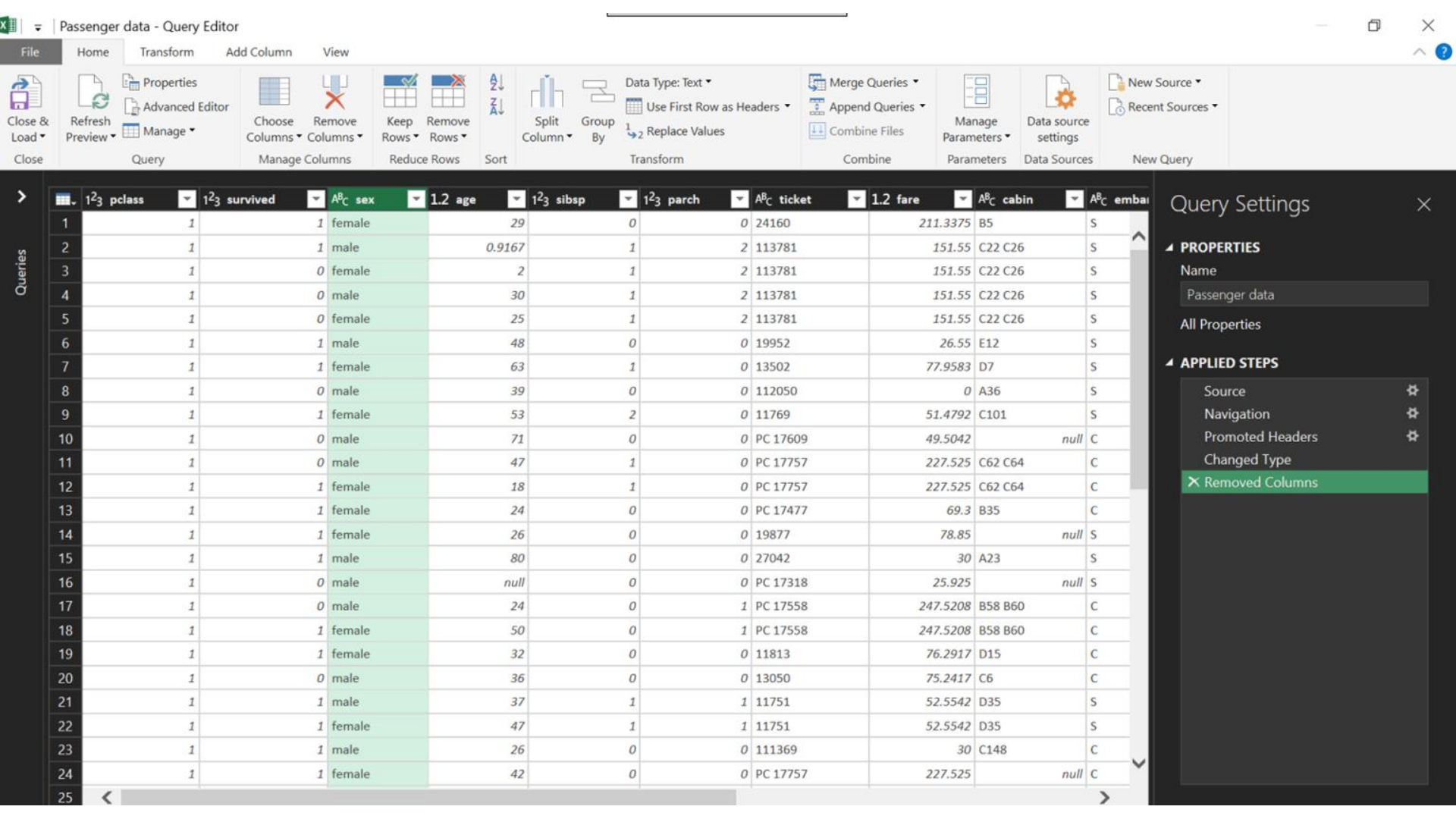
- Data is never clean – it always contains missing values, errors, incorrect formats, and other problems that make it impossible to feed to a machine learning model without preprocessing.
- This is what data cleansing is all about – correcting all these problems before starting the real analysis.
- As an example of how to clean a dataset, we will use the Titanic passengers dataset.


File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text to Columns
What-If Analysis
Forecast Sheet
Group
Ungroup
Subtotal
Data Analysis
Sort & Filter
Data Tools
Forecast
Outline
Analysis
From File
From Database
From Azure
From Online Services
From Other Sources
Combine Queries
From Workbook
From Text/CSV
From XML
From JSON
From Folder
From SharePoint Folder
Launch Query Editor...
Data Catalog Search
My Data Catalog Queries
Data Source Settings...
Query Options


Navigator
Select multiple items
Display Options
titanic.xls [2]
Dictionary
Passenger data
Passenger data
pclass survived name sex
1 1 Allen, Miss. Elisabeth Walton female
1 1 Allison, Master. Hudson Trevor male
1 0 Allison, Miss. Helen Loraine female
1 0 Allison, Mr. Hudson Joshua Creighton male
1 0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
1 1 Anderson, Mr. Harry male
1 1 Andrews, Miss. Kornelia Theodosia female
1 0 Andrews, Mr. Thomas Jr male
1 1 Appleton, Mrs. Edward Dale (Charlotte Lamson) female
1 0 Artagaveytia, Mr. Ramon male
1 0 Astor, Col. John Jacob male
1 1 Astor, Mrs. John Jacob (Madeleine Talmadge Force) female
1 1 Aubart, Mme. Leontine Pauline female
1 1 Barber, Miss. Ellen "Nellie" female
1 1 Barkworth, Mr. Algernon Henry Wilson male
1 0 Baumann, Mr. John D male
1 0 Baxter, Mr. Quigg Edmond male
1 1 Baxter, Mrs. James (Helene DeLaudeniere Chaput) female
1 1 Bazzani, Miss. Albina female
1 0 Beattie, Mr. Thomson male
1 1 Beckwith, Mr. Richard Leonard male
1 1 Beckwith, Mrs. Richard Leonard (Sallie Monypeny) female
1 1 Behr, Mr. Karl Howell male
Load
Edit
Cancel

Cleansing data

- Click on Edit and start the data cleansing process.
- The first thing we notice is that we don't need the column containing the passenger names; it gives no useful information to our analysis.
- In fact, in most cases, we will be required to remove personal information from our data, due to privacy policies.
- Select the name column.


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
pclass
survived
sex
age
sibsp
parch
ticket
fare
cabin
embar
1 1 female 29 0 0 24160 211.3375 B5 S
2 1 1 male 0.9167 1 2 113781 151.55 C22 C26 S
3 1 0 female 2 1 2 113781 151.55 C22 C26 S
4 1 0 male 30 1 2 113781 151.55 C22 C26 S
5 1 0 female 25 1 2 113781 151.55 C22 C26 S
6 1 1 male 48 0 0 19952 26.55 E12 S
7 1 1 female 63 1 0 13502 77.9583 D7 S
8 1 0 male 39 0 0 112050 0 A36 S
9 1 1 female 53 2 0 11769 51.4792 C101 S
10 1 0 male 71 0 0 PC 17609 49.5042 null C
11 1 0 male 47 1 0 PC 17757 227.525 C62 C64 C
12 1 1 female 18 1 0 PC 17757 227.525 C62 C64 C
13 1 1 female 24 0 0 PC 17477 69.3 B35 C
14 1 1 female 26 0 0 19877 78.85 null S
15 1 1 male 80 0 0 27042 30 A23 S
16 1 0 male null 0 0 PC 17318 25.925 null S
17 1 0 male 24 0 1 PC 17558 247.5208 B58 B60 C
18 1 1 female 50 0 1 PC 17558 247.5208 B58 B60 C
19 1 1 female 32 0 0 11813 76.2917 D15 C
20 1 0 male 36 0 0 13050 75.2417 C6 C
21 1 1 male 37 1 1 11751 52.5542 D35 S
22 1 1 female 47 1 1 11751 52.5542 D35 S
23 1 1 male 26 0 0 111369 30 C148 C
24 1 1 female 42 0 0 PC 17757 227.525 null C
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns

Passenger data - Query Editor

File | Home | Transform | Add Column | View

Close & Load | Refresh Preview | Properties | Advanced Editor | Manage | Close | Query

Choose Columns | Remove Columns | Manage Columns

Keep Rows | Remove Rows | Reduce Rows

Sort

Split Column | Group By | Data Type: Text | Use First Row as Headers | Replace Values | Transform

Merge Queries | Append Queries | Combine Files | Combine

Manage Parameters | Parameters

Data source settings | Data Sources

New Source | Recent Sources | New Query

Queries

	pclass	survived	sex	age	sibsp	parch	ticket	fare	cabin	embar
1	1	1	female	29	0	0	24160	211.3375	B5	S
2	1	1	male	0.9167	1	2	113781	151.55	C22 C26	S
3	1	0	female	2	1	2	113781	151.55	C22 C26	S
4	1	0	male	30	1	2	113781	151.55	C22 C26	S
5	1	0	female	25	1	2	113781	151.55	C22 C26	S
6	1	1	male	48	0	0	19952	26.55	E12	S
7	1	1	female	63	1	0	13502	77.9583	D7	S
8	1	0	male	39	0	0	112050	0	A36	S
9	1	1	female	53	2	0	11769	51.4792	C101	S
10	1	0	male	71	0	0	PC 17609	49.5042	unknown	C
11	1	0	male	47	1	0	PC 17757	227.525	C62 C64	C
12	1	1	female	18	1	0	PC 17757	227.525	C62 C64	C
13	1	1	female	24	0	0	PC 17477	69.3	B35	C
14	1	1	female	26	0	0	19877	78.85	unknown	S
15	1	1	male	80	0	0	27042	30	A23	S
16	1	0	male	null	0	0	PC 17318	25.925	unknown	S
17	1	0	male	24	0	1	PC 17558	247.5208	B58 B60	C
18	1	1	female	50	0	1	PC 17558	247.5208	B58 B60	C
19	1	1	female	32	0	0	11813	76.2917	D15	C
20	1	0	male	36	0	0	13050	75.2417	C6	C
21	1	1	male	37	1	1	11751	52.5542	D35	S
22	1	1	female	47	1	1	11751	52.5542	D35	S
23	1	1	male	26	0	0	111369	30	C148	C
24	1	1	female	42	0	0	PC 17757	227.525	unknown	C
25										

Query Settings

PROPERTIES

Name

Passenger data

All Properties

APPLIED STEPS

- Source
- Navigation
- Promoted Headers
- Changed Type
- Removed Columns
- Replaced Value

13 COLUMNS, 999+ ROWS

PREVIEW DOWNLOADED AT 12:21

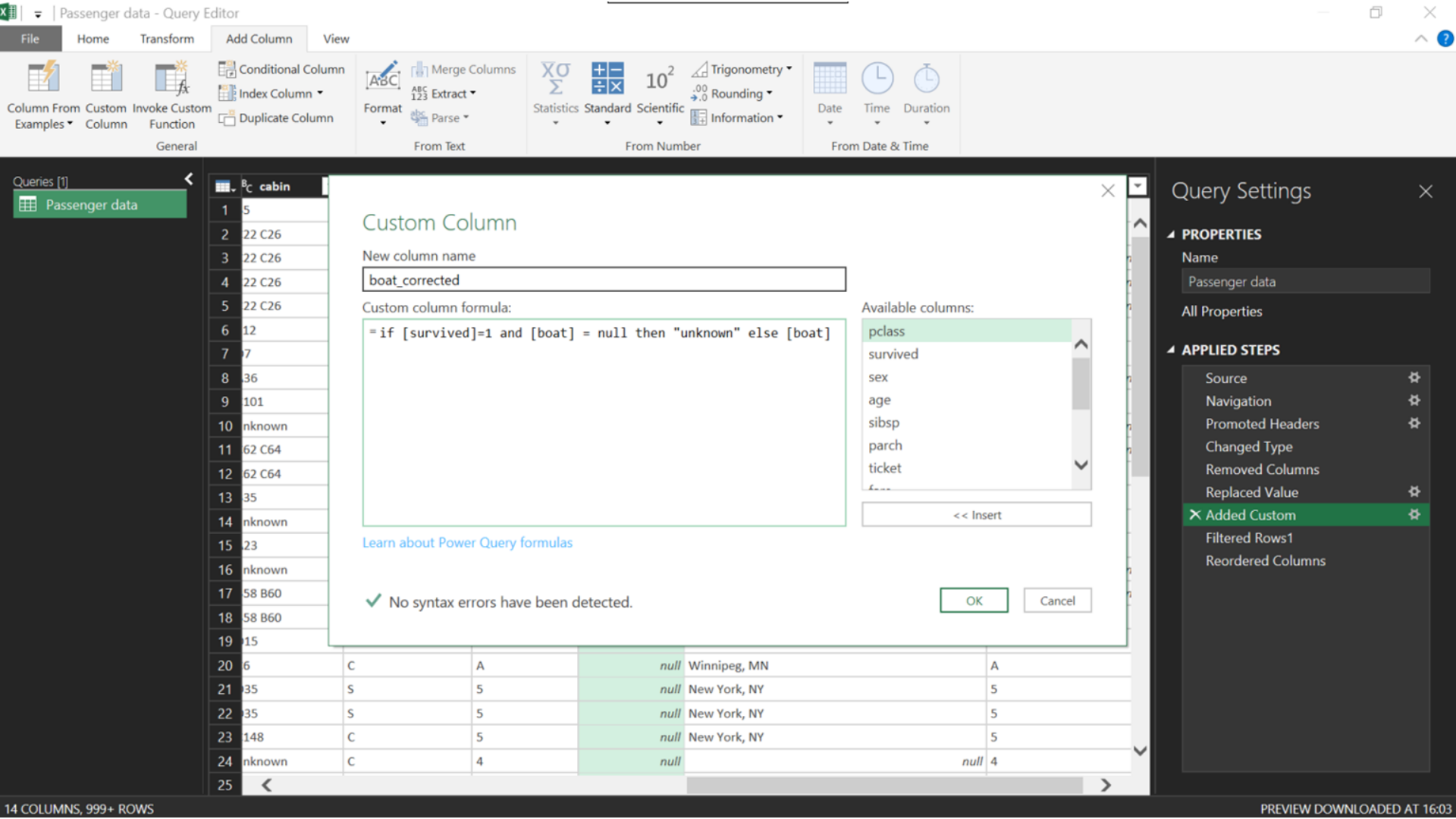
Cleansing data

- In Query Editor, select the Add Column tab.
- Select Custom Column.
- The dialog shows us an option to name the new column and define its contents.
- Type boat_corrected into the textbox.

Cleansing data

- Define a function to calculate the column's contents, as follows:

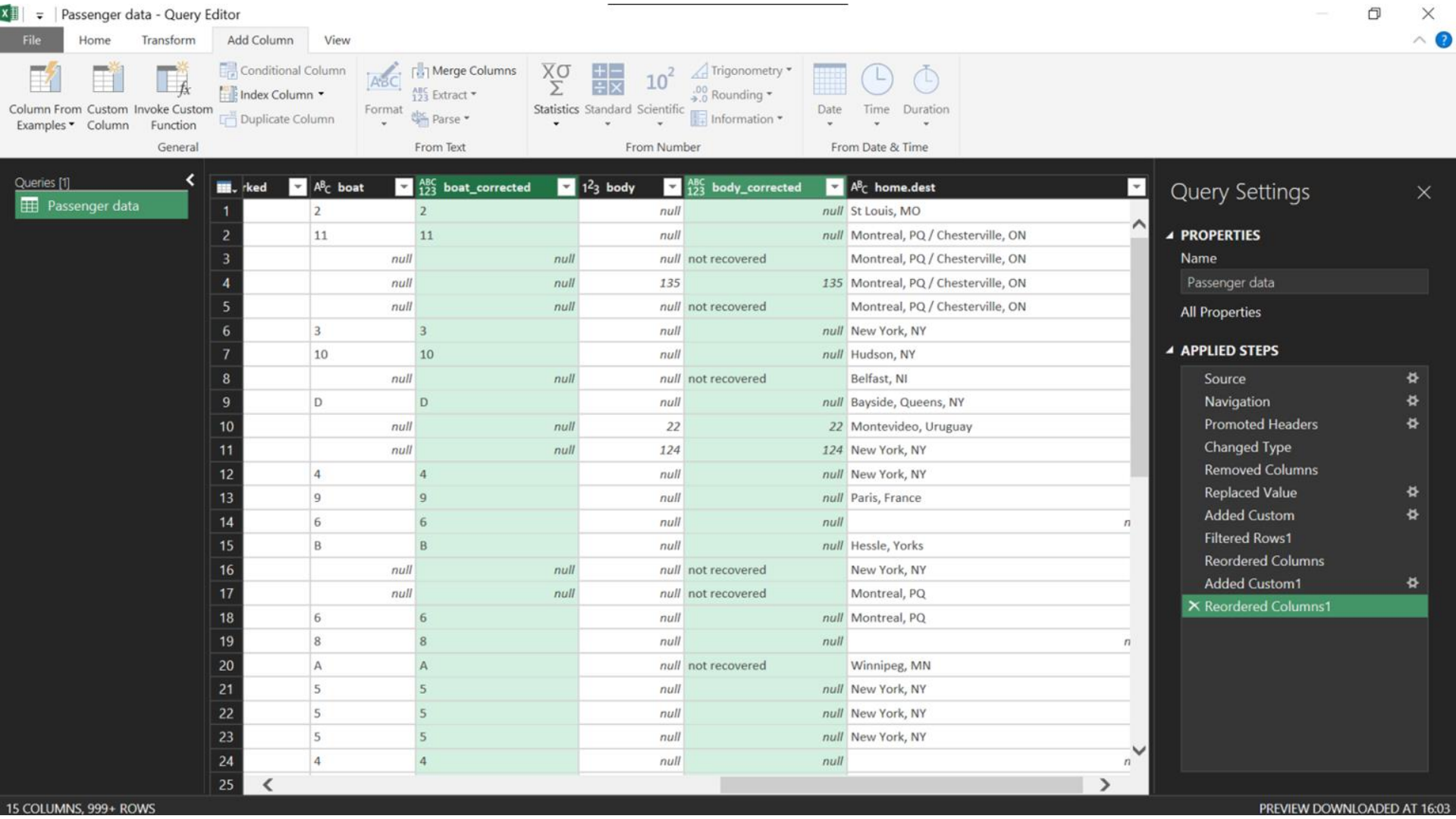
```
if [survived]=1 and [boat] = null then "unknown" else [boat]
```

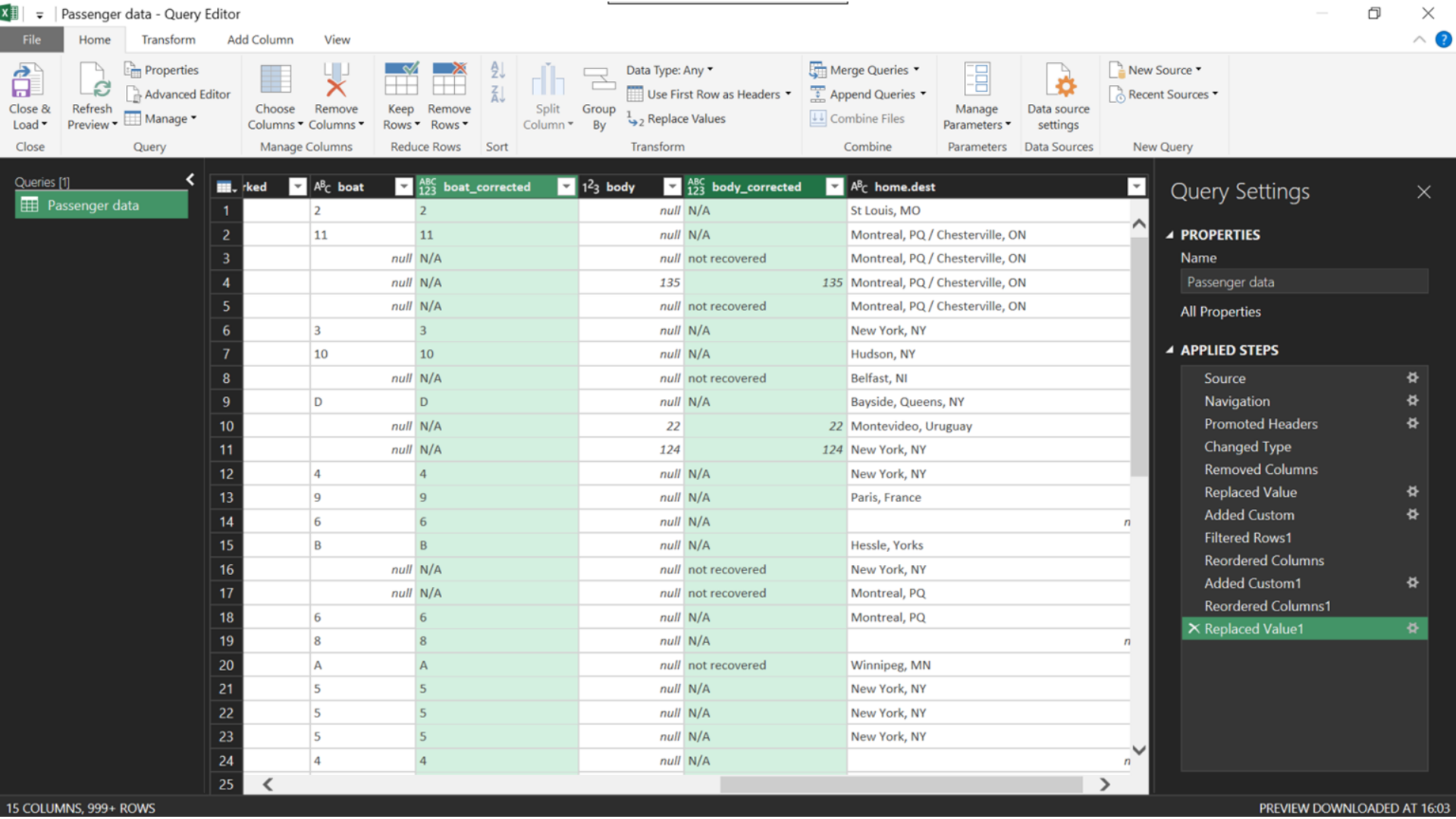

Passenger data - Query Editor
File
Home
Transform
Add Column
View
Column From Examples
Custom Column
Invoke Custom Function
Conditional Column
Index Column
Duplicate Column
General
Format
Merge Columns
Extract
Parse
From Text
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
From Number
Date
Time
Duration
From Date & Time
Queries [1]
Passenger data
cabin
Custom Column
New column name
boat_corrected
Custom column formula:
=if [survived]=1 and [boat] = null then "unknown" else [boat]
Available columns:
pclass
survived
sex
age
sibsp
parch
ticket
<< Insert
Learn about Power Query formulas
No syntax errors have been detected.
OK
Cancel
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
14 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03

Cleansing data

- Add another new column in order to correct the values in body and define a different value for the column:

if [survived]=0 and [body] = null then "not recovered" else [body]


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Column From Examples
Custom Column
Invoke Custom Function
Conditional Column
Index Column
Duplicate Column
General
Format
Merge Columns
Extract
Parse
From Text
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
From Number
Date
Time
Duration
From Date & Time
Queries [1]
Passenger data
rked
boat
boat_corrected
body
body_corrected
home.dest
1 2 2 null null St Louis, MO
2 11 11 null null Montreal, PQ / Chesterville, ON
3 null null null not recovered Montreal, PQ / Chesterville, ON
4 null null 135 135 Montreal, PQ / Chesterville, ON
5 null null null not recovered Montreal, PQ / Chesterville, ON
6 3 3 null null New York, NY
7 10 10 null null Hudson, NY
8 null null null not recovered Belfast, NI
9 D D null null Bayside, Queens, NY
10 null null 22 22 Montevideo, Uruguay
11 null null 124 124 New York, NY
12 4 4 null null New York, NY
13 9 9 null null Paris, France
14 6 6 null null n
15 B B null null Hessle, Yorks
16 null null null not recovered New York, NY
17 null null null not recovered Montreal, PQ
18 6 6 null null Montreal, PQ
19 8 8 null null n
20 A A null not recovered Winnipeg, MN
21 5 5 null null New York, NY
22 5 5 null null New York, NY
23 5 5 null null New York, NY
24 4 4 null null n
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
15 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Any
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries [1]
Passenger data
rked
boat
boat_corrected
body
body_corrected
home.dest
1 2 2 null N/A St Louis, MO
2 11 11 null N/A Montreal, PQ / Chesterville, ON
3 null N/A null not recovered Montreal, PQ / Chesterville, ON
4 null N/A 135 135 Montreal, PQ / Chesterville, ON
5 null N/A null not recovered Montreal, PQ / Chesterville, ON
6 3 3 null N/A New York, NY
7 10 10 null N/A Hudson, NY
8 null N/A null not recovered Belfast, NI
9 D D null N/A Bayside, Queens, NY
10 null N/A 22 22 Montevideo, Uruguay
11 null N/A 124 124 New York, NY
12 4 4 null N/A New York, NY
13 9 9 null N/A Paris, France
14 6 6 null N/A n
15 B B null N/A Hessle, Yorks
16 null N/A null not recovered New York, NY
17 null N/A null not recovered Montreal, PQ
18 6 6 null N/A Montreal, PQ
19 8 8 null N/A n
20 A A null not recovered Winnipeg, MN
21 5 5 null N/A New York, NY
22 5 5 null N/A New York, NY
23 5 5 null N/A New York, NY
24 4 4 null N/A n
25
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
Replaced Value1
15 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 16:03

Cleansing data

- Replace all the missing values (null) with -1. We can do this easily by selecting the column and clicking on Replace Values.

- Navigate to the Add Column tab.

Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name

Age group

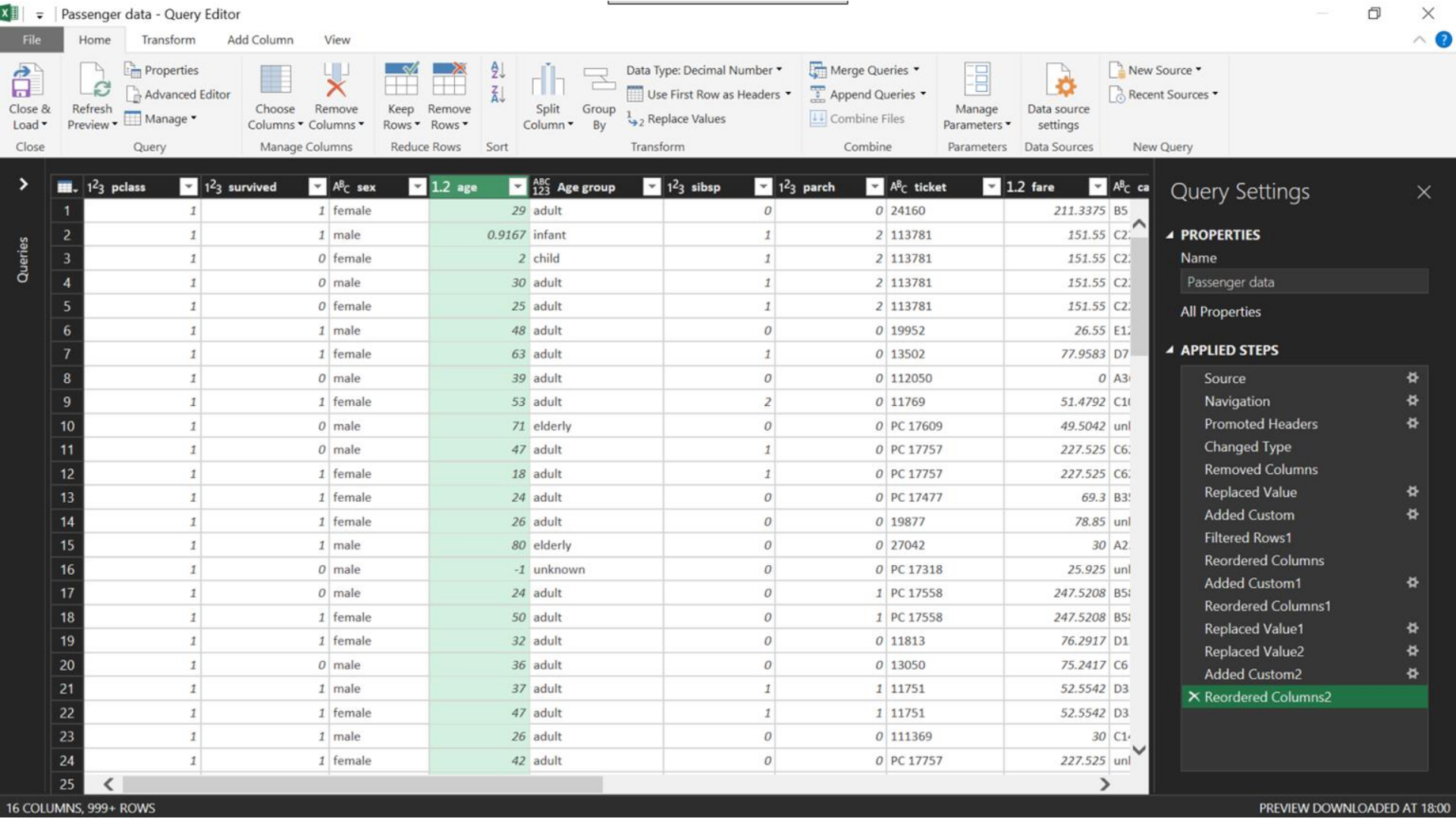
	Column Name	Operator	Value		Output	
If	age	equals	-1	Then	unknown	...
Else If	age	is less than	1	Then	infant	
Else If	age	is less than	12	Then	child	
Else If	age	is less than	18	Then	teenager	
Else If	age	is less than	65	Then	adult	
Else If	age	is greater than or...	65	Then	elderly	

Add rule

Otherwise

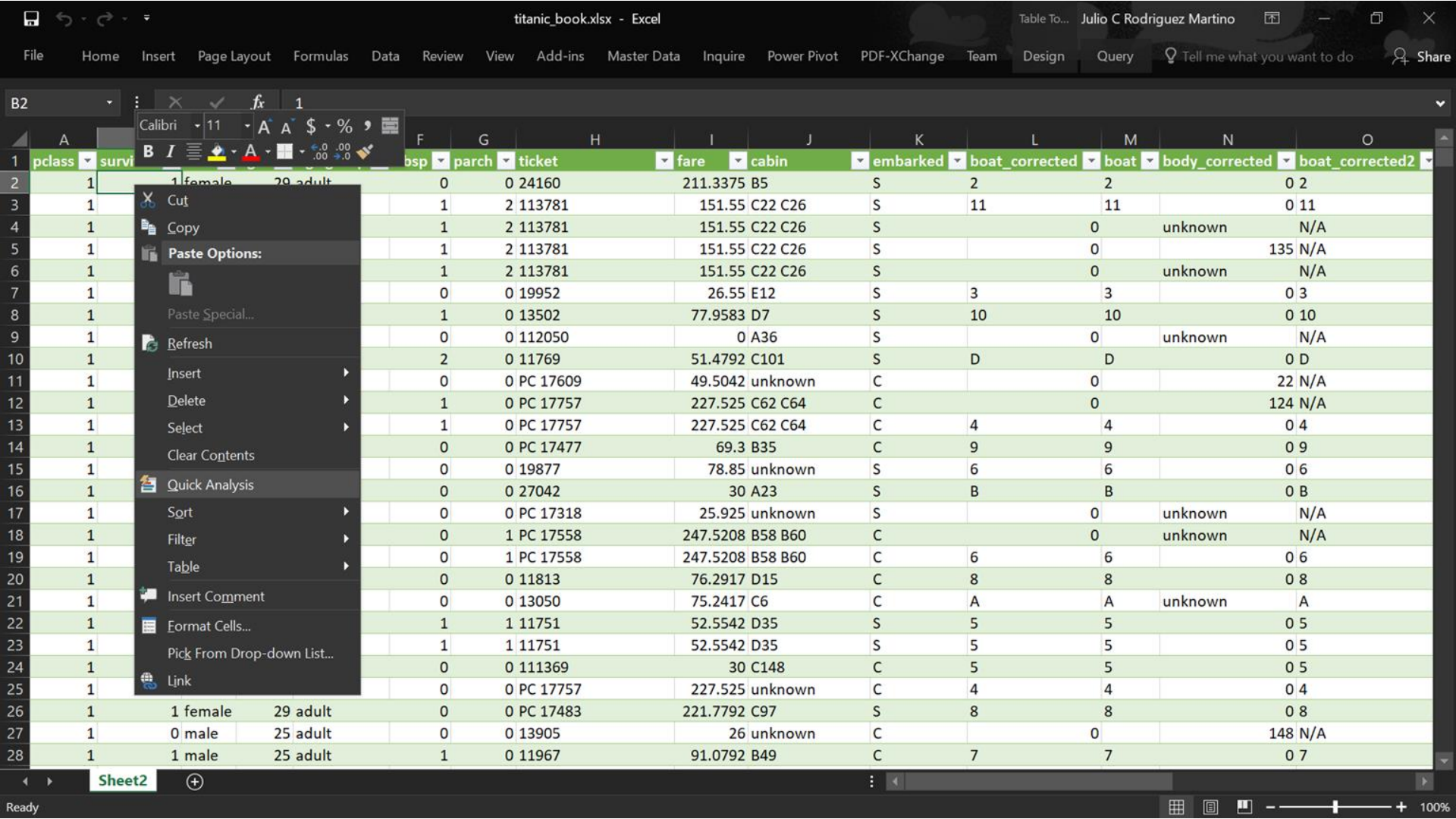
unknown

OK Cancel


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Decimal Number
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Manage Parameters
Data source settings
New Source
Recent Sources
Close
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
Parameters
Data Sources
New Query
Queries
pclass
survived
sex
age
Age group
sibsp
parch
ticket
fare
1 female 29 adult 0 0 24160 211.3375 B5
1 male 0.9167 infant 1 2 113781 151.55 C2
1 female 2 child 1 2 113781 151.55 C2
1 male 30 adult 1 2 113781 151.55 C2
1 female 25 adult 1 2 113781 151.55 C2
1 male 48 adult 0 0 19952 26.55 E1
1 female 63 adult 1 0 13502 77.9583 D7
1 male 39 adult 0 0 112050 0 A3
1 female 53 adult 2 0 11769 51.4792 C1
1 male 71 elderly 0 0 PC 17609 49.5042 unl
1 male 47 adult 1 0 PC 17757 227.525 C6
1 female 18 adult 1 0 PC 17757 227.525 C6
1 female 24 adult 0 0 PC 17477 69.3 B3
1 female 26 adult 0 0 19877 78.85 unl
1 male 80 elderly 0 0 27042 30 A2
1 male -1 unknown 0 0 PC 17318 25.925 unl
1 male 24 adult 0 1 PC 17558 247.5208 B5
1 female 50 adult 0 1 PC 17558 247.5208 B5
1 female 32 adult 0 0 11813 76.2917 D1
1 male 36 adult 0 0 13050 75.2417 C6
1 male 37 adult 1 1 11751 52.5542 D3
1 female 47 adult 1 1 11751 52.5542 D3
1 male 26 adult 0 0 111369 30 C1
1 female 42 adult 0 0 PC 17757 227.525 unl
Query Settings
PROPERTIES
Name
Passenger data
All Properties
APPLIED STEPS
Source
Navigation
Promoted Headers
Changed Type
Removed Columns
Replaced Value
Added Custom
Filtered Rows1
Reordered Columns
Added Custom1
Reordered Columns1
Replaced Value1
Replaced Value2
Added Custom2
Reordered Columns2
16 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 18:00

Visualizing data for preliminary analysis

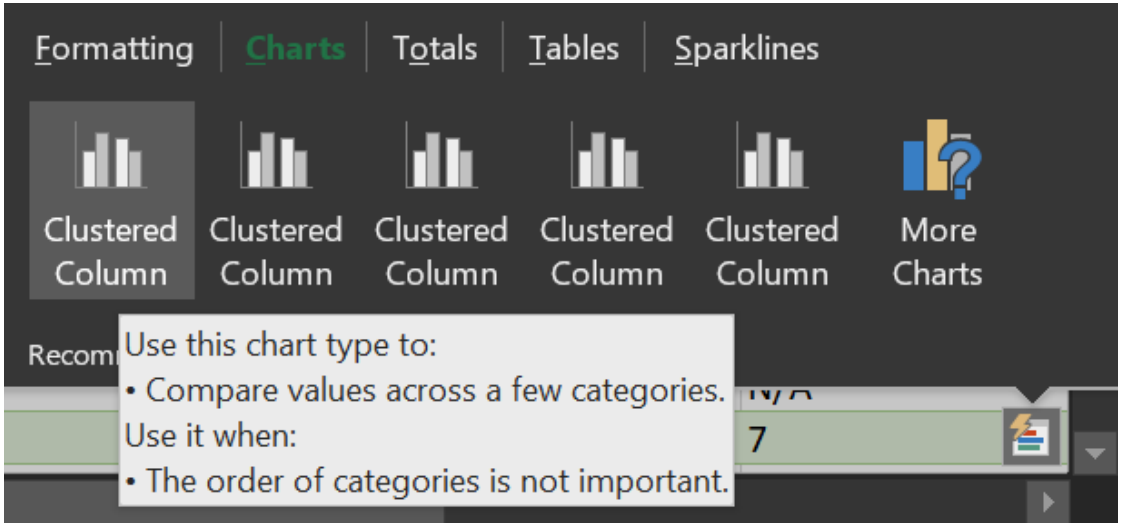
- After cleaning the dataset, it is always recommended to visualize it.
- This helps us gain an understanding of the different variables, how their values are distributed, and the correlations that exist between them (we will explore correlations in more detail in the next lesson).
- We can determine which variables are important to our analyses, which ones give us more information, and which ones can be discarded for being redundant.



pclass	survi			bsp	parch	ticket	fare	cabin	embarked	boat_corrected	boat	body_corrected	boat_corrected2
1	1	female	29 adult	0	0	24160	211.3375	B5	S	2	2	0	2
1				1	2	113781	151.55	C22 C26	S	11	11	0	11
1				1	2	113781	151.55	C22 C26	S		0	unknown	N/A
1				1	2	113781	151.55	C22 C26	S		0	135	N/A
1				1	2	113781	151.55	C22 C26	S		0	unknown	N/A
1				0	0	19952	26.55	E12	S	3	3	0	3
1				1	0	13502	77.9583	D7	S	10	10	0	10
1				0	0	112050	0	A36	S		0	unknown	N/A
1				2	0	11769	51.4792	C101	S	D	D	0	D
1				0	0	PC 17609	49.5042	unknown	C		0	22	N/A
1				1	0	PC 17757	227.525	C62 C64	C		0	124	N/A
1				1	0	PC 17757	227.525	C62 C64	C	4	4	0	4
1				0	0	PC 17477	69.3	B35	C	9	9	0	9
1				0	0	19877	78.85	unknown	S	6	6	0	6
1				0	0	27042	30	A23	S	B	B	0	B
1				0	0	PC 17318	25.925	unknown	S		0	unknown	N/A
1				0	1	PC 17558	247.5208	B58 B60	C		0	unknown	N/A
1				0	1	PC 17558	247.5208	B58 B60	C	6	6	0	6
1				0	0	11813	76.2917	D15	C	8	8	0	8
1				0	0	13050	75.2417	C6	C	A	A	unknown	A
1				1	1	11751	52.5542	D35	S	5	5	0	5
1				1	1	11751	52.5542	D35	S	5	5	0	5
1				0	0	111369	30	C148	C	5	5	0	5
1				0	0	PC 17757	227.525	unknown	C	4	4	0	4
1	1	female	29 adult	0	0	PC 17483	221.7792	C97	S	8	8	0	8
1	0	male	25 adult	0	0	13905	26	unknown	C		0	148	N/A
1	1	male	25 adult	1	0	11967	91.0792	B49	C	7	7	0	7

Visualizing data for preliminary analysis

- In the pop-up window, we can choose the chart type. Select Clustered Column, as shown in the following screenshot:



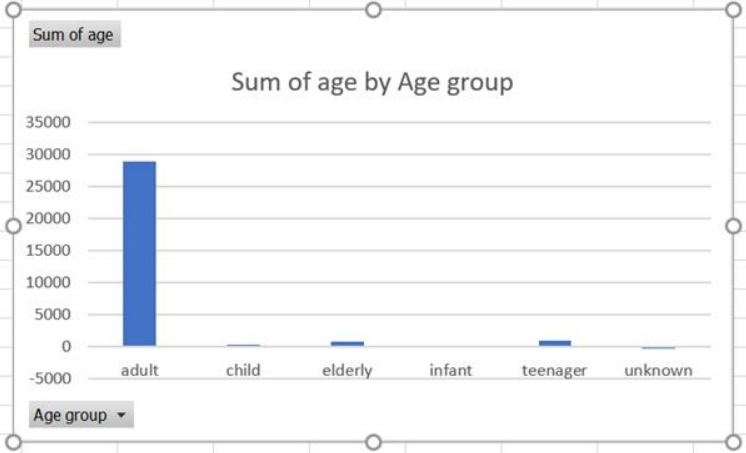
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	Sum of age
adult	28943.5
child	417.5
elderly	910.5
infant	8.1667
teenager	976
unknown	-263



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Move Up
Move Down
Move to Beginning
Move to End
Move to Report Filter
Move to Axis Fields (Categories)
Move to Legend Fields (Series)
Move to Values
Hide Value Field Buttons on Chart
Hide All Field Buttons on Chart
Remove Field
Value Field Settings...

Drag fields between areas b

Filters

Axis (Categories)

Age group

Sum of age

Defer Layout Update

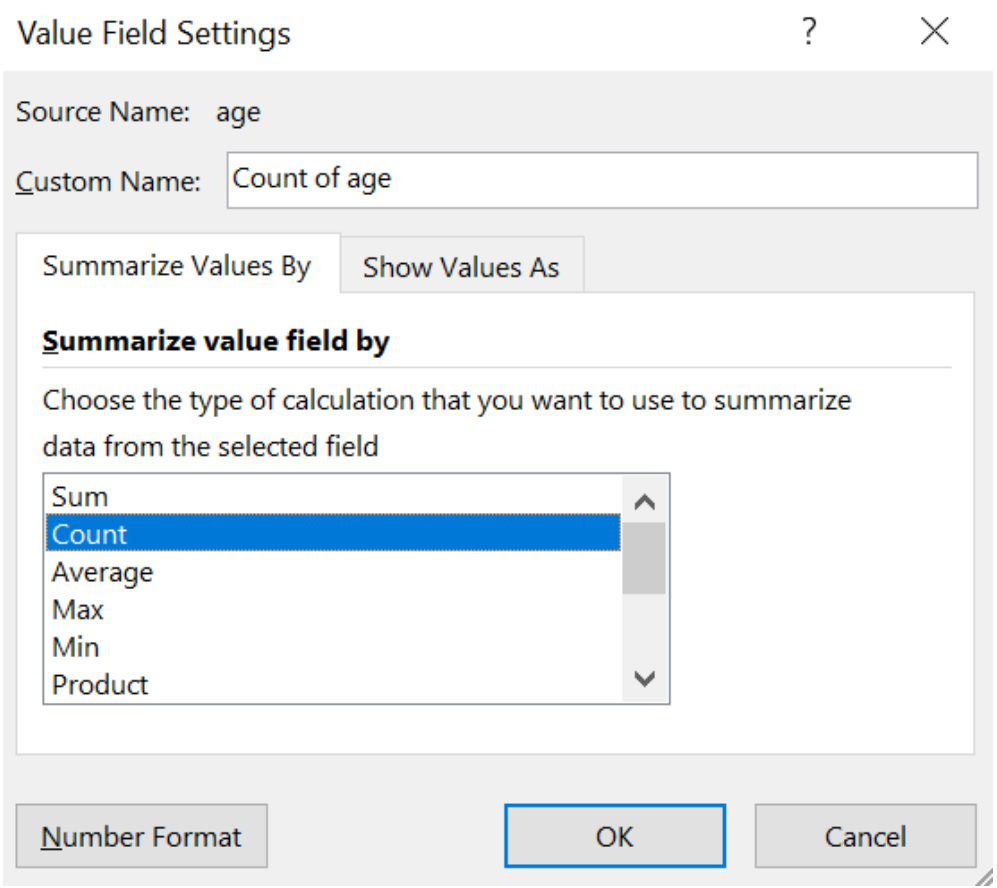
Update

Sheet1 Sheet2

Ready Calculate

100%

- Click on Value Field Settings; you will see a pop-up window, similar to the one in the following screenshot, where you can change from Sum to Count, since we want to Count the values, and then calculate the Sum of them:



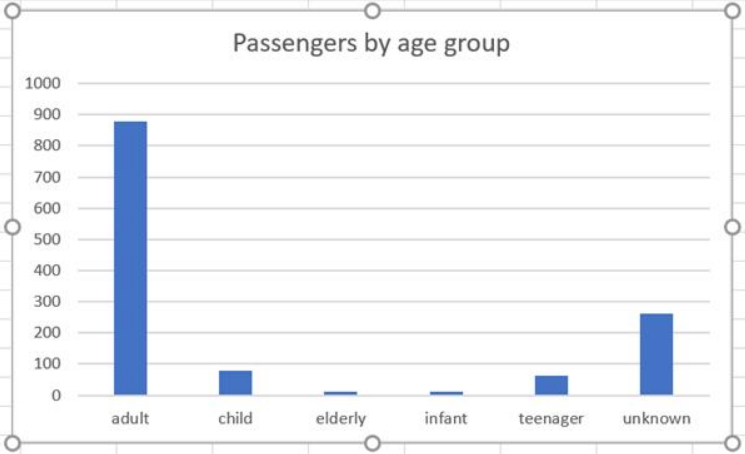
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	Count of Age group
adult	879
child	79
elderly	13
infant	12
teenager	63
unknown	263



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): Age group

Values: Count of Age group

Defer Layout Update | Update

Sheet1 Sheet2

Ready

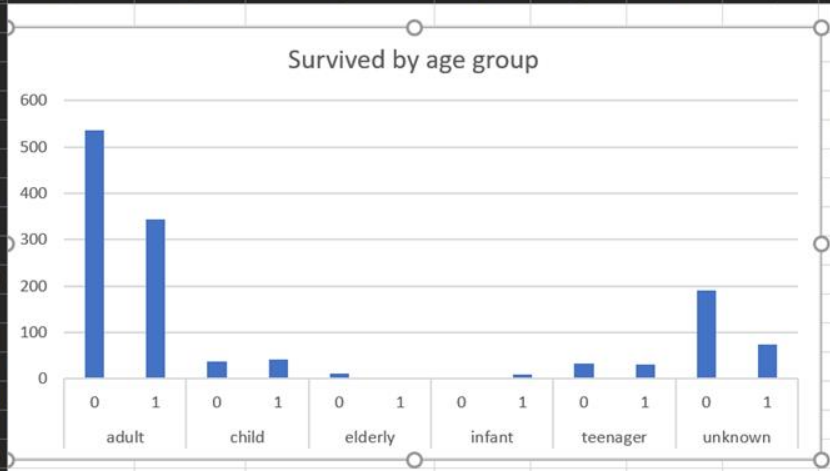
100%

titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- **survived**
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

Age group

survived

Values

Count of Age group

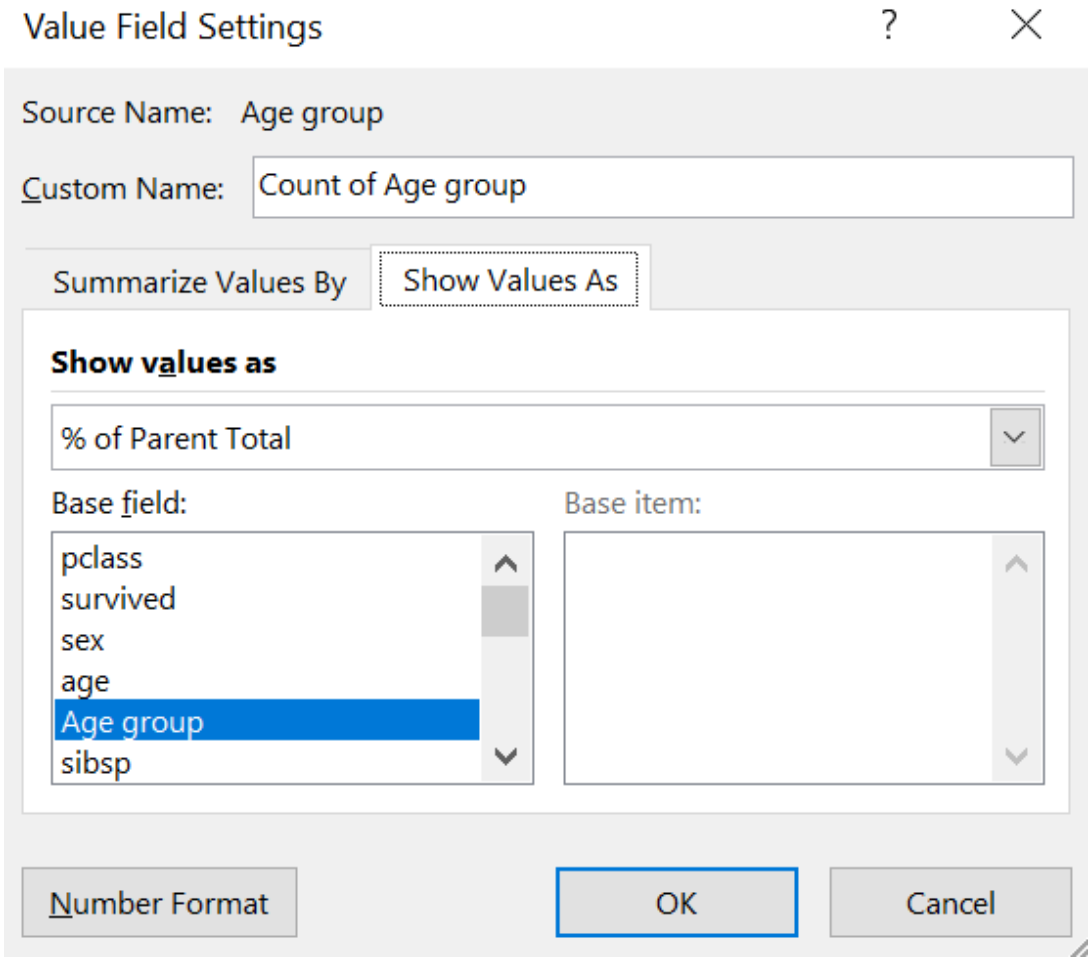
Defer Layout Update

Update

Sheet1 Sheet2

Ready

100%


Value Field Settings
Source Name: Age group
Custom Name: Count of Age group
Summarize Values By
Show Values As
Show values as
% of Parent Total
Base field:
pclass
survived
sex
age
Age group
sibsp
Base item:
Number Format
OK
Cancel

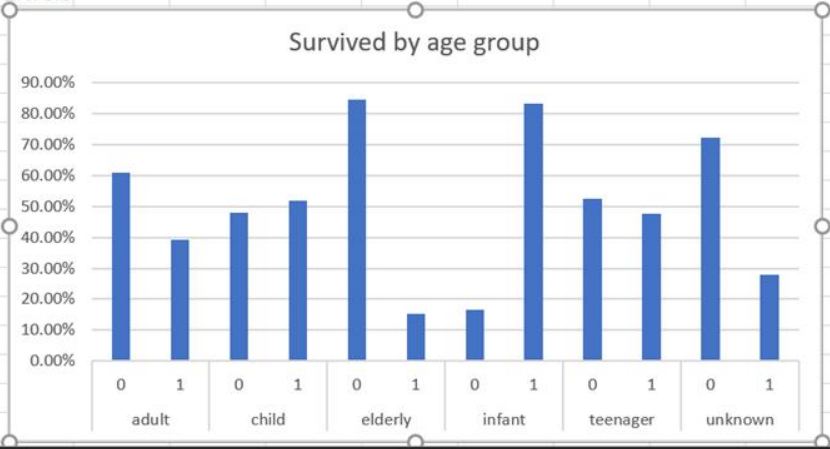
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

Age group	survived	Count of Age group
adult	0	60.86%
adult	1	39.14%
child	0	48.10%
child	1	51.90%
elderly	0	84.62%
elderly	1	15.38%
infant	0	16.67%
infant	1	83.33%
teenager	0	52.38%
teenager	1	47.62%
unknown	0	72.24%
unknown	1	27.76%



Sheet1 Sheet2

PivotChart Fields

Choose fields to add to report:

Search

- pclass
- **survived**
- sex
- age
- **Age group**
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters | Legend (Series)

Axis (Categories): Age group, survived

Values: Count of Age group

Defer Layout Update Update

Ready 100%

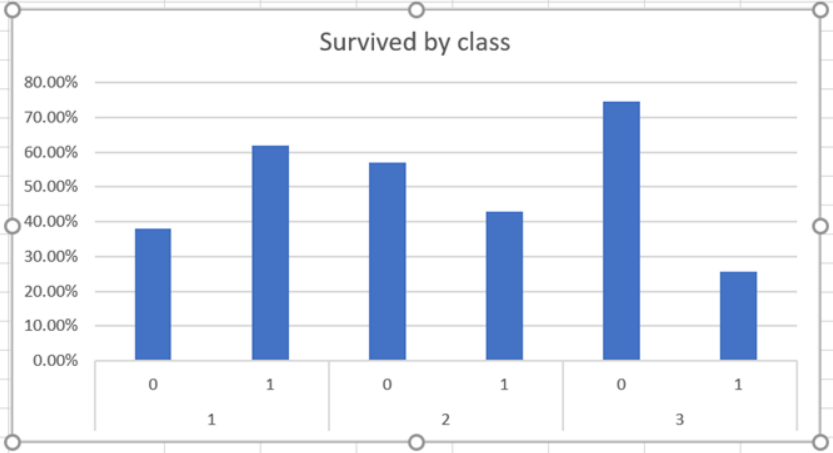
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

pclass	survived	Sum of pclass
1	0	38.08%
1	1	61.92%
2	0	57.04%
2	1	42.96%
3	0	74.47%
3	1	25.53%



Sheet1 Sheet2

Ready

PivotChart Fields

Choose fields to add to report:

Search

- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare
- cabin

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

pclass

survived

Values

Sum of pclass

Defer Layout Update

Update

100%

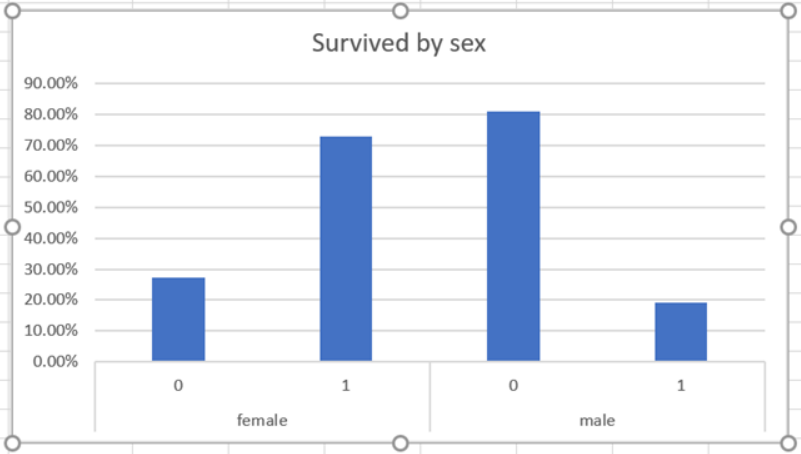
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

sex	survived	Count of sex
fema	0	27.25%
fema	1	72.75%
male	0	80.90%
male	1	19.10%



PivotChart Fields

Choose fields to add to report:

Search

- boat_corrected
- boat
- body_corrected
- boat_corrected2
- body
- body_corrected3
- home.dest
- % Total

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

sex

survived

Values

Count of sex

Defer Layout Update

Update

Sheet1 Sheet2

ESP 09:28

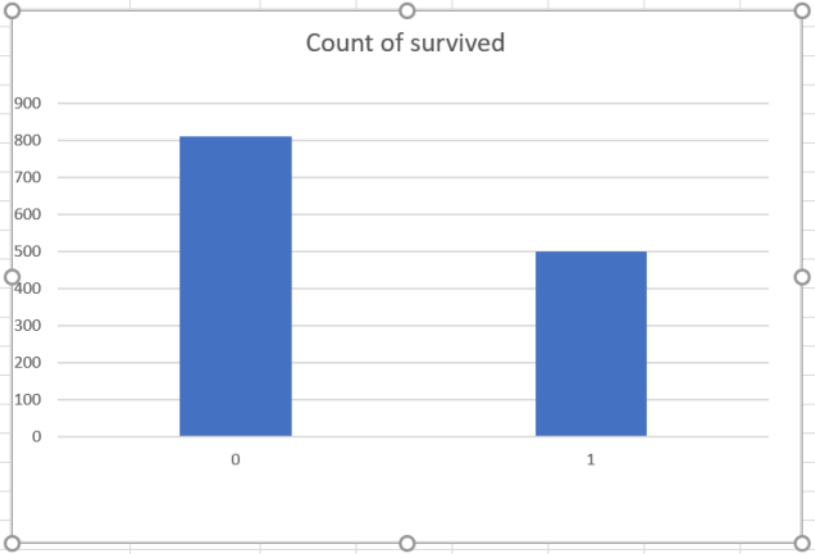
titanic_book.xlsx - Excel

Julio C Rodriguez Martino

File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Analyze Design Format Tell me Share

Chart 1

survived	Count of survived
0	809
1	500



PivotChart Fields

Choose fields to add to report:

Search

- pclass
- survived
- sex
- age
- Age group
- sibsp
- parch
- ticket
- fare

Drag fields between areas below:

Filters

Legend (Series)

Axis (Categories)

survived

Values

Count of survived

Defer Layout Update

Update

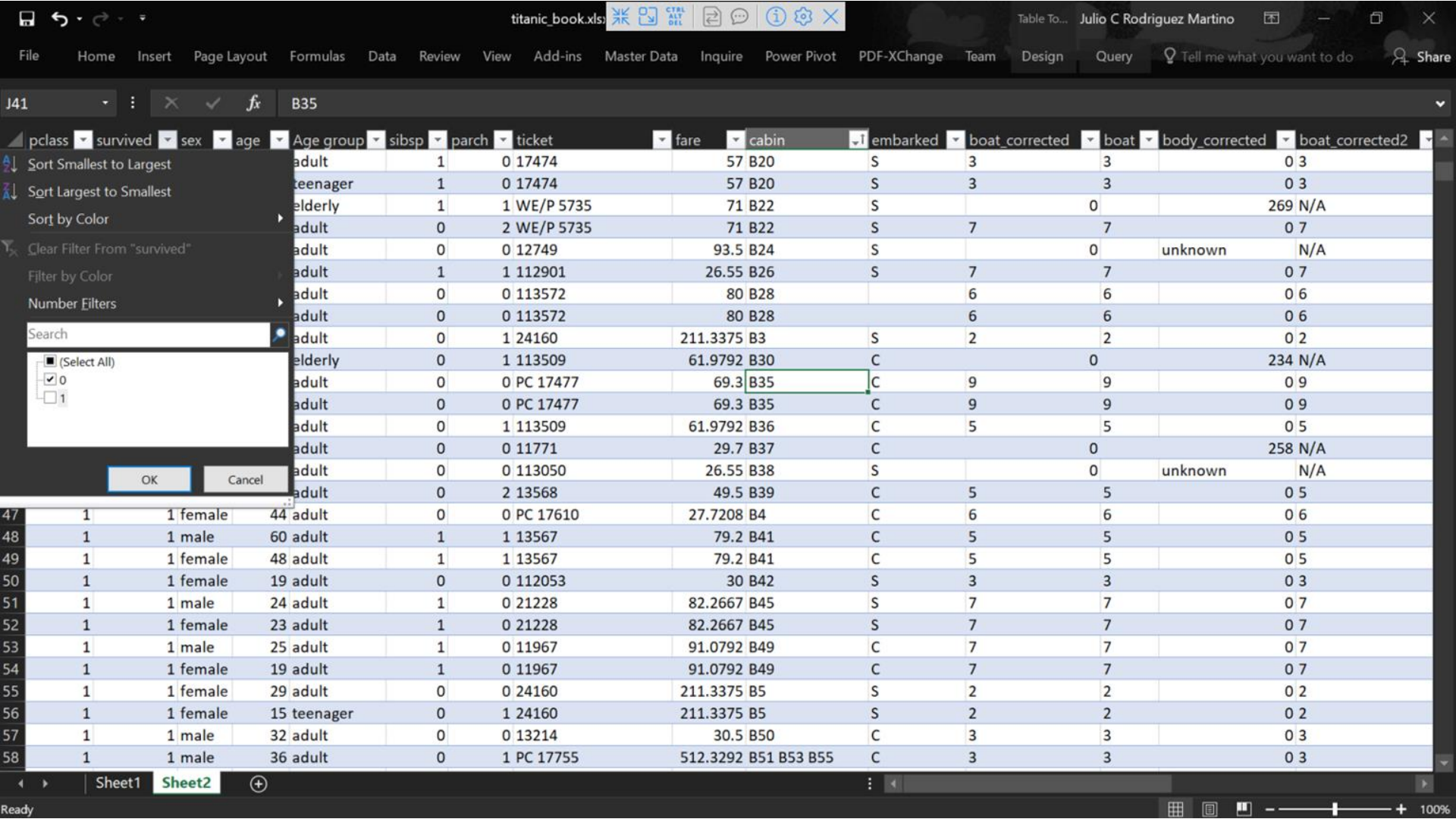
Sheet1 Sheet2

Ready

100%

Understanding unbalanced datasets

- It is clear from the preceding diagram and table that there are nearly twice as many non-survivors than survivors.
- If we use this dataset as is, we are introducing a bias to our dataset that will affect the results.
- Predicting 0 for the survival variable will be approximately two times more probable than predicting 1.

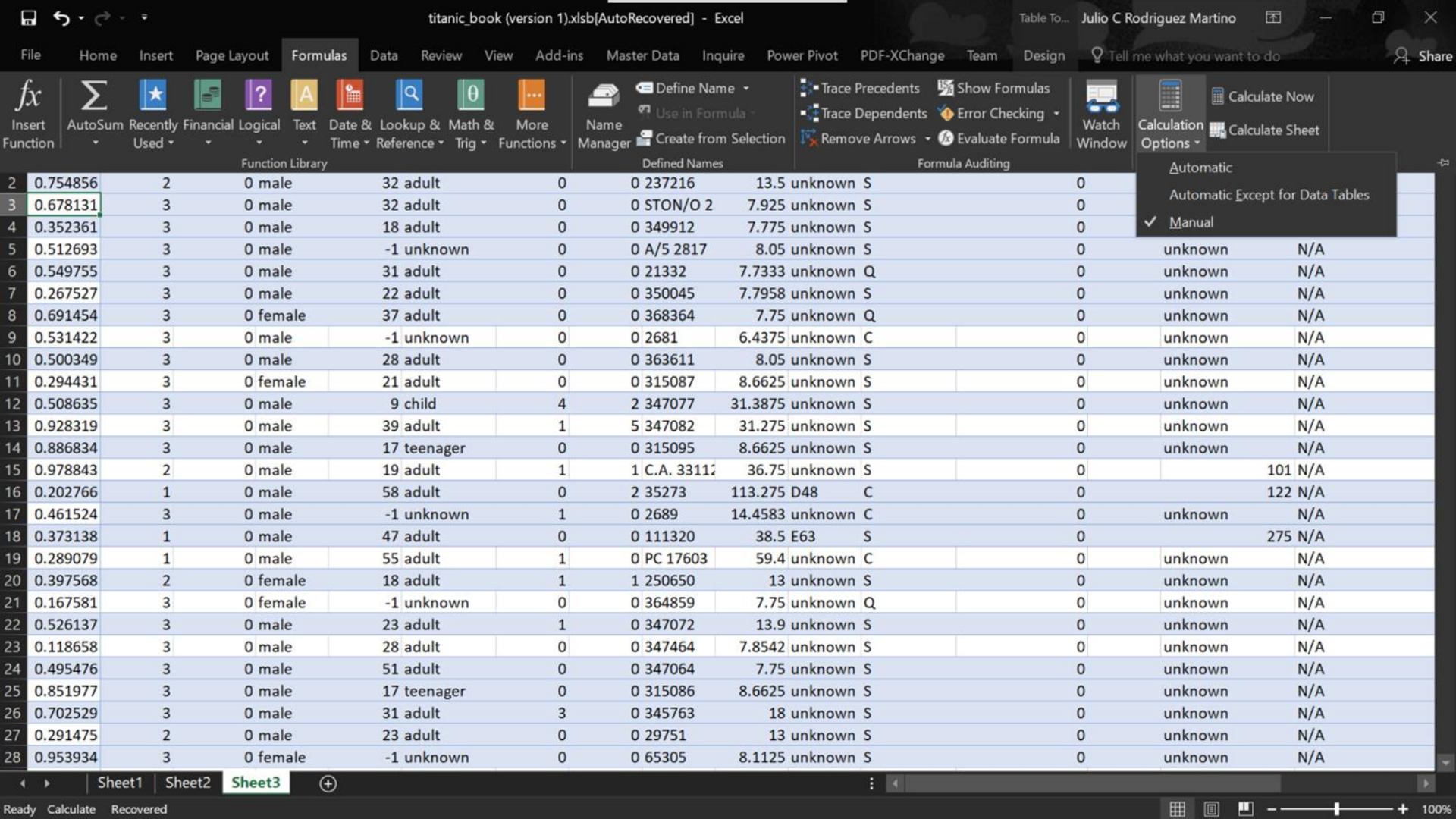


pclass	survived	sex	age	Age group	sibsp	parch	ticket	fare	cabin	embarked	boat_corrected	boat	body_corrected	boat_corrected2
				adult	1	0	17474	57	B20	S	3	3	0	3
				teenager	1	0	17474	57	B20	S	3	3	0	3
				elderly	1	1	WE/P 5735	71	B22	S		0	269	N/A
				adult	0	2	WE/P 5735	71	B22	S	7	7	0	7
				adult	0	0	12749	93.5	B24	S		0	unknown	N/A
				adult	1	1	112901	26.55	B26	S	7	7	0	7
				adult	0	0	113572	80	B28		6	6	0	6
				adult	0	0	113572	80	B28		6	6	0	6
				adult	0	1	24160	211.3375	B3	S	2	2	0	2
				elderly	0	1	113509	61.9792	B30	C		0	234	N/A
				adult	0	0	PC 17477	69.3	B35	C	9	9	0	9
				adult	0	0	PC 17477	69.3	B35	C	9	9	0	9
				adult	0	1	113509	61.9792	B36	C	5	5	0	5
				adult	0	0	11771	29.7	B37	C		0	258	N/A
				adult	0	0	113050	26.55	B38	S		0	unknown	N/A
				adult	0	2	13568	49.5	B39	C	5	5	0	5
47	1	1	female	44	adult	0	0	PC 17610	27.7208	B4	C	6	6	0	6
48	1	1	male	60	adult	1	1	13567	79.2	B41	C	5	5	0	5
49	1	1	female	48	adult	1	1	13567	79.2	B41	C	5	5	0	5
50	1	1	female	19	adult	0	0	112053	30	B42	S	3	3	0	3
51	1	1	male	24	adult	1	0	21228	82.2667	B45	S	7	7	0	7
52	1	1	female	23	adult	1	0	21228	82.2667	B45	S	7	7	0	7
53	1	1	male	25	adult	1	0	11967	91.0792	B49	C	7	7	0	7
54	1	1	female	19	adult	1	0	11967	91.0792	B49	C	7	7	0	7
55	1	1	female	29	adult	0	0	24160	211.3375	B5	S	2	2	0	2
56	1	1	female	15	teenager	0	1	24160	211.3375	B5	S	2	2	0	2
57	1	1	male	32	adult	0	0	13214	30.5	B50	C	3	3	0	3
58	1	1	male	36	adult	0	1	PC 17755	512.3292	B51 B53 B55	C	3	3	0	3

Understanding unbalanced datasets

- Copy the entries and paste them into a new worksheet.
- Insert a new column at the beginning, named ID.
- Turn the data into a table (Insert | Table, keeping the first row as headers).
- Enter the following formula in the first cell and copy it into the rest of the column:

=RAND()



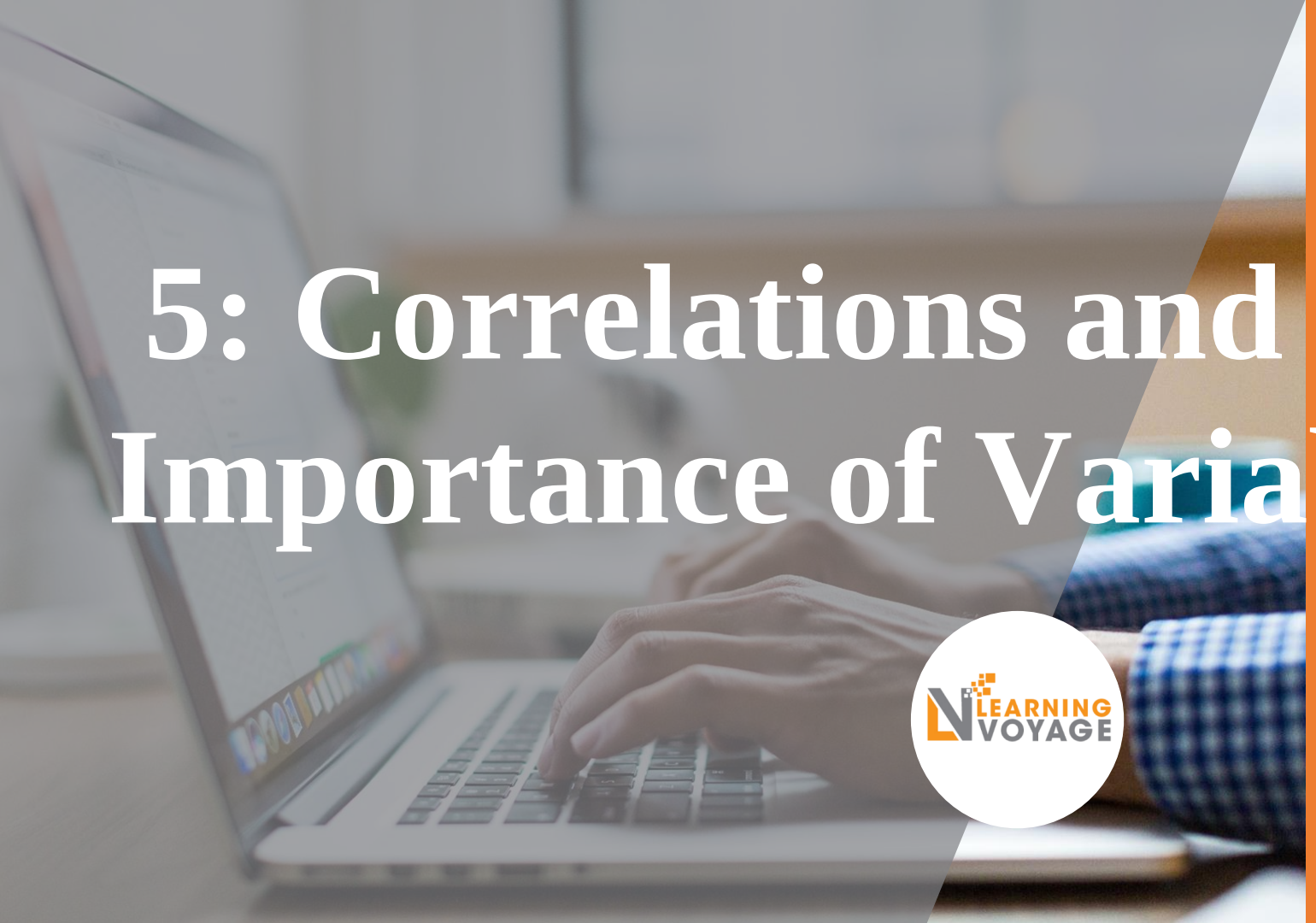
2	0.754856	2	0 male	32 adult	0	0	237216	13.5	unknown	S		0		
3	0.678131	3	0 male	32 adult	0	0	STON/O 2	7.925	unknown	S		0		
4	0.352361	3	0 male	18 adult	0	0	349912	7.775	unknown	S		0		
5	0.512693	3	0 male	-1 unknown	0	0	A/5 2817	8.05	unknown	S		0	unknown	N/A
6	0.549755	3	0 male	31 adult	0	0	21332	7.7333	unknown	Q		0	unknown	N/A
7	0.267527	3	0 male	22 adult	0	0	350045	7.7958	unknown	S		0	unknown	N/A
8	0.691454	3	0 female	37 adult	0	0	368364	7.75	unknown	Q		0	unknown	N/A
9	0.531422	3	0 male	-1 unknown	0	0	2681	6.4375	unknown	C		0	unknown	N/A
10	0.500349	3	0 male	28 adult	0	0	363611	8.05	unknown	S		0	unknown	N/A
11	0.294431	3	0 female	21 adult	0	0	315087	8.6625	unknown	S		0	unknown	N/A
12	0.508635	3	0 male	9 child	4	2	347077	31.3875	unknown	S		0	unknown	N/A
13	0.928319	3	0 male	39 adult	1	5	347082	31.275	unknown	S		0	unknown	N/A
14	0.886834	3	0 male	17 teenager	0	0	315095	8.6625	unknown	S		0	unknown	N/A
15	0.978843	2	0 male	19 adult	1	1	C.A. 3311	36.75	unknown	S		0	101	N/A
16	0.202766	1	0 male	58 adult	0	2	35273	113.275	D48	C		0	122	N/A
17	0.461524	3	0 male	-1 unknown	1	0	2689	14.4583	unknown	C		0	unknown	N/A
18	0.373138	1	0 male	47 adult	0	0	111320	38.5	E63	S		0	275	N/A
19	0.289079	1	0 male	55 adult	1	0	PC 17603	59.4	unknown	C		0	unknown	N/A
20	0.397568	2	0 female	18 adult	1	1	250650	13	unknown	S		0	unknown	N/A
21	0.167581	3	0 female	-1 unknown	0	0	364859	7.75	unknown	Q		0	unknown	N/A
22	0.526137	3	0 male	23 adult	1	0	347072	13.9	unknown	S		0	unknown	N/A
23	0.118658	3	0 male	28 adult	0	0	347464	7.8542	unknown	S		0	unknown	N/A
24	0.495476	3	0 male	51 adult	0	0	347064	7.75	unknown	S		0	unknown	N/A
25	0.851977	3	0 male	17 teenager	0	0	315086	8.6625	unknown	S		0	unknown	N/A
26	0.702529	3	0 male	31 adult	3	0	345763	18	unknown	S		0	unknown	N/A
27	0.291475	2	0 male	23 adult	0	0	29751	13	unknown	S		0	unknown	N/A
28	0.953934	3	0 female	-1 unknown	0	0	65305	8.1125	unknown	S		0	unknown	N/A

Understanding unbalanced datasets

- Order the data by ID (you can choose ascending or descending order, it does not make any difference).
- Select the first 500 rows to be your random sample.
- Copy these rows to a new sheet.
- Add the 500 rows with survived as 1.

Summary

- In this lesson, we explored different methods of dealing with missing data and learned how to group or summarize it.
- We have shown you how important it is to visualize the data after cleaning, in order to be able to understand and interpret the results, from basic to more advanced model predictions.
- This is the beginning of any feature engineering, since we transform and/or discard features based on their values



5: Correlations and the Importance of Variables

LEARNING VOYAGE

Correlations and the Importance of Variables

- Correlation between variables, in general, means that a change in one variable reflects on the other.
- However, it does not mean that the change in one variable is caused by the change in the correlated variable.
- For example, the selling price of a product is correlated to its manufacturing cost, but the price increase is not totally caused by it, since there are other factors such as transportation and inflation to take into account.

Correlations and the Importance of Variables

In this lesson, we will cover the following topics:

- Building a scatter diagram
- Calculating the covariance
- Calculating the Pearson's coefficient of correlation
- Studying the Spearman's correlation
- Understanding least squares
- Focusing on feature selection

Technical requirements

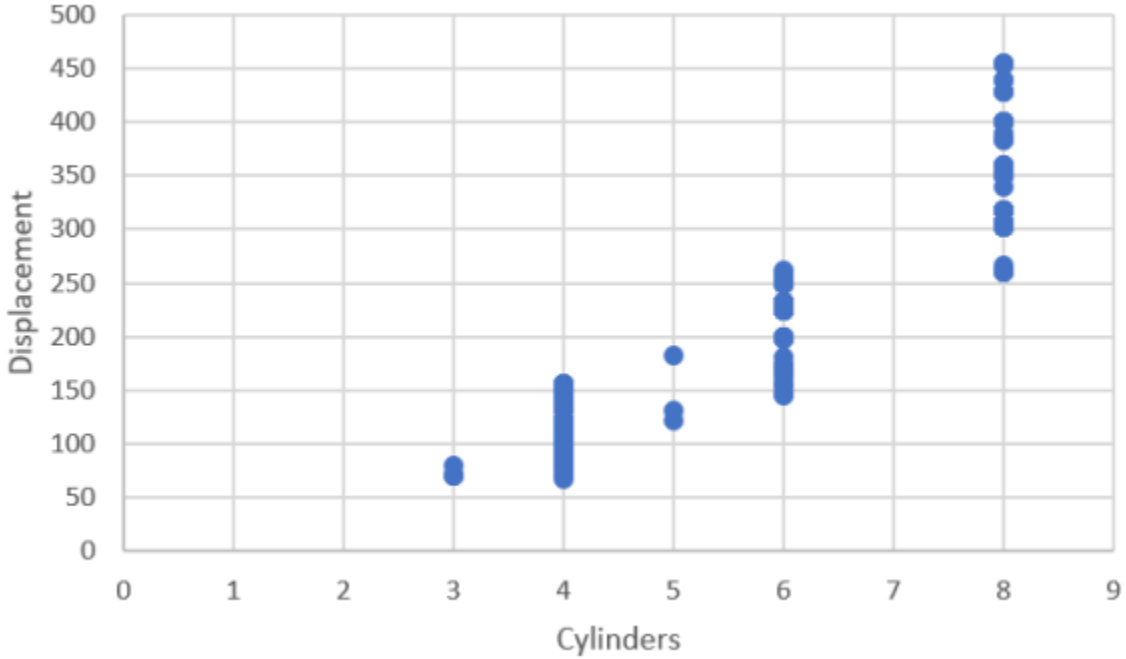
- You will need to download the auto-mpg.xlsx file from the GitHub.

Building a scatter diagram

- First, load the auto-mpg.xlsx file. We will use the data in it to illustrate different aspects of this lesson.
- The meaning of the variables are described in the Excel file and in its references.
- The simplest way of assessing correlations between variables is to create a scatter diagram, taking all features in pairs.

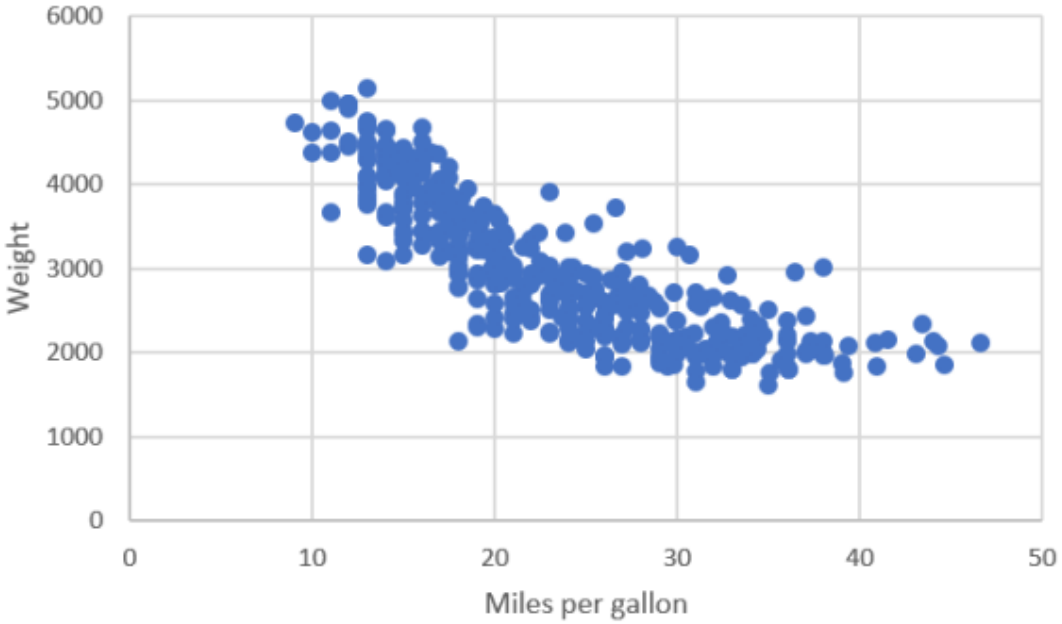
Building a scatter diagram

- The scatter diagram can be seen in the following diagram:

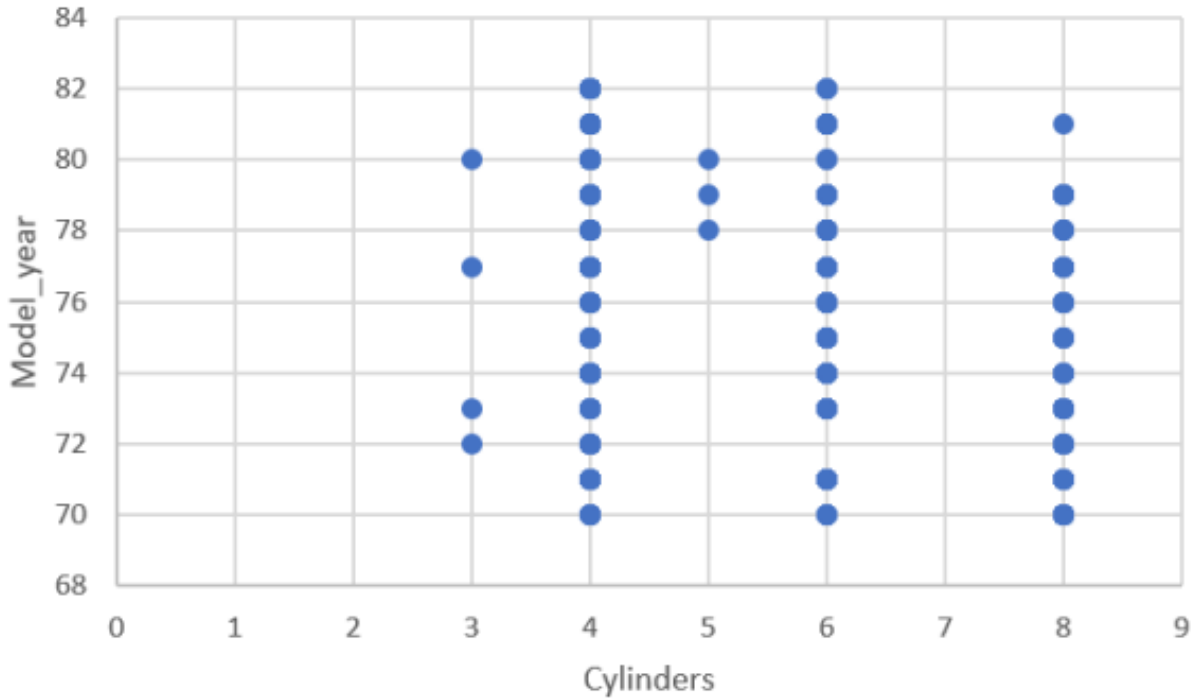


Building a scatter diagram

- If we, instead, look at the relationship between fuel consumption and car weight, the diagram will be similar to the following:



Building a scatter diagram



Building a scatter diagram

- This method of finding correlations in scatter diagrams is fine if we have a few variables, but the number of diagrams needs scales fast.

- In fact, if the number of variables is Nv, then the number of combinations needed to see all correlations is as follows:

Nv * (Nv -1)

Calculating the covariance

- We need to define a statistical method that quantitatively measures the degree of association between two features.
- The covariance of two variables does exactly that, so let's see how it is calculated. If there are two variables, x and y, we first center their values around their mean values, and ; then, we multiply the new values and take the mean of the product:

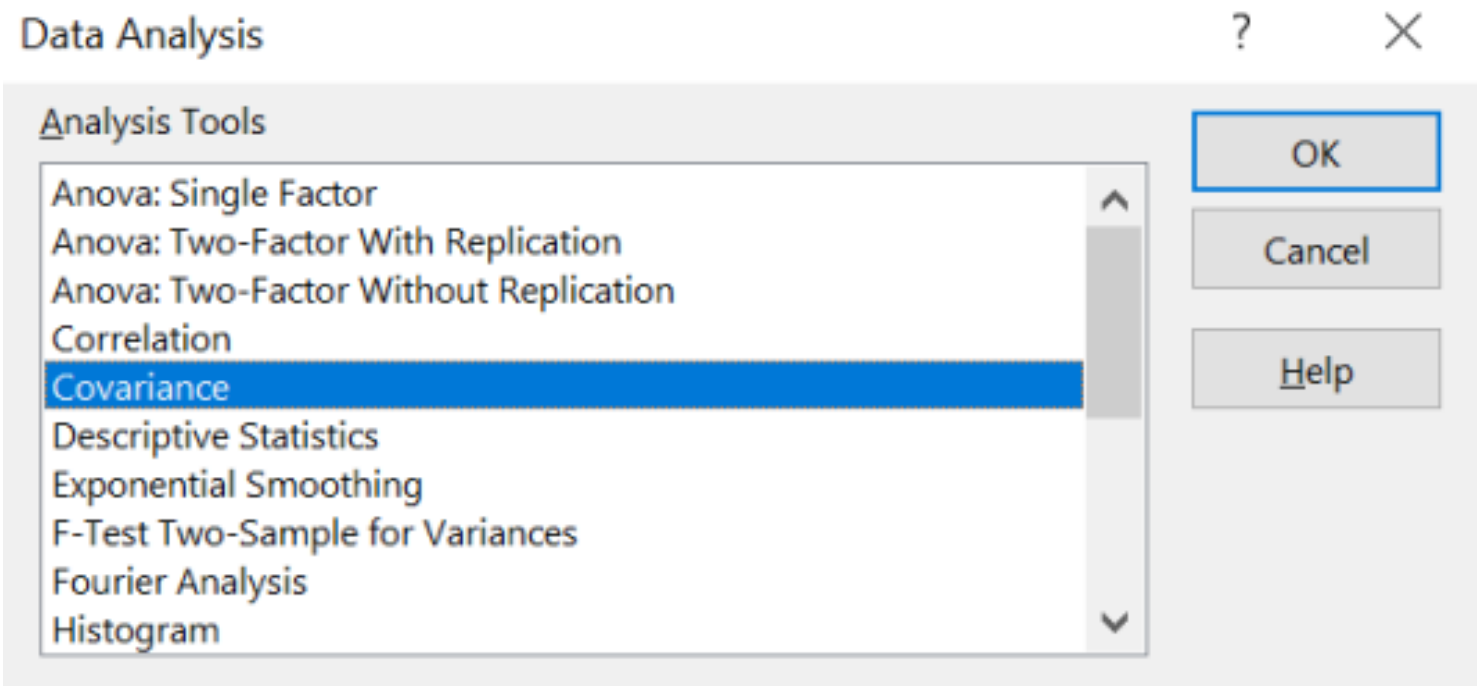
$$Cov(x, y) = mean[(x - \hat{x}).(y - \hat{y})]$$

Calculating the covariance

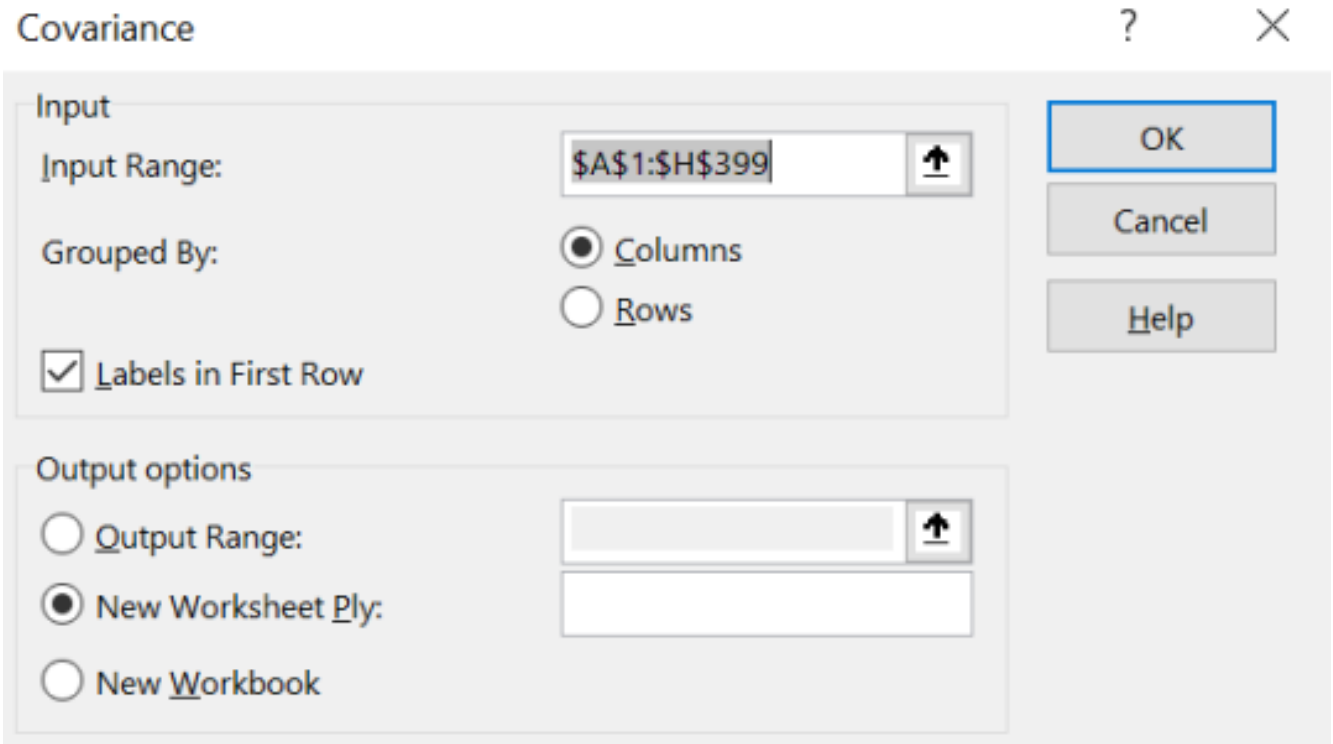
To calculate the covariances, perform the following steps:

- Open the data file.
- Navigate to Data | Data Analysis.
- In the pop-up window, select Covariance, as shown in the next slide.

Calculating the covariance



Calculating the covariance



Calculating the covariance

- The result is the following table:

	A	B	C	D	E	F	G	H	I
1		*mpg*	*cylinders*	*displacement*	*horsepower*	*weight*	*acceleration*	*model_year*	*origin*
2	mpg	60.93611929							
3	cylinders	-10.28300927	2.886146						
4	displacem	-653.7555781	168.1995	10844.88207					
5	horsepow	-233.2613494	55.20705	3604.81427	1477.789879				
6	weight	-5491.379555	1287.453	82161.4674	28193.51406	715339.1287			
7	acceleratic	9.036168531	-2.36489	-155.9401792	-73.00026551	-972.4495158	7.585740575		
8	model_ye	16.69909977	-2.18799	-142.3585516	-58.88582882	-957.5344183	2.930722709	13.63808995	
9	origin	3.523310017	-0.76555	-50.83693594	-14.07673886	-393.647774	0.45420949	0.534443575	0.641676
10									

Calculating the Pearson's coefficient of correlation

- The Pearson's coefficient is most commonly used when comparing two variables and it works by measuring the linear relationship between them.

- The original definition given by Pearson is as follows:

$$\rho_{x,y} = \frac{\sum_i (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_i (x_i - \hat{x})^2 (y_i - \hat{y})^2}}$$

Calculating the Pearson's coefficient of correlation

- The resulting table is as follows:

	A	B	C	D	E	F	G	H	I
1		*mpg*	*cylinders*	*displacement*	*horsepower*	*weight*	*acceleration*	*model_year*	*origin*
2	mpg	1							
3	cylinders	-0.7754	1						
4	displacement	-0.8042	0.950721	1					
5	horsepower	-0.77843	0.842983	0.897257002	1				
6	weight	-0.83174	0.896017	0.932824147	0.864537738	1			
7	acceleration	0.420289	-0.50542	-0.543684084	-0.68919551	-0.41745732	1		
8	model_year	0.579267	-0.34875	-0.370164161	-0.416361477	-0.306564334	0.288136954	1	
9	origin	0.56345	-0.56254	-0.609409399	-0.455171453	-0.581023914	0.205873007	0.180662195	1
10									

Calculating the Pearson's coefficient of correlation

- Another definition for the Pearson coefficient is as follows:

$$\rho_{x,y} = b. \frac{\sigma_x}{\sigma_y}$$

Studying the Spearman's correlation

- To calculate the Spearman's coefficient, we need to first rank the values of each variable, that is, the order of the values when we sort them from highest to lowest.
- Once we have the new table, we will calculate Pearson's ρ on it.
- In a new sheet, we define the following formula in a cell:

=RANK.AVG(Data!A2;auto_mpg[mpg])

	A	B	C	D	E	F	G	H
1	Rank_mpg	Rank_cylinders	Rank_displacement	Rank_horsepower	Rank_weight	Rank_acceleration	Rank_model_year	Rank_origin
2	283	52	75	94	109	362.5	384	274
3	337.5	52	46.5	35.5	90	372	384	274
4	283	52	65	56.5	115	384	384	274
5	318	52	84	56.5	116	362.5	384	274
6	303	52	93	81	112	388	384	274
7	337.5	52	8	12.5	34	390.5	384	274
8	356	52	4	5	33	395	384	274
9	356	52	5.5	7	37	396.5	384	274
10	356	52	2	3	25	390.5	384	274
11	337.5	52	23	16	76	396.5	384	274
12	337.5	52	24.5	30	104	390.5	384	274
13	356	52	56	38.5	100	398	384	274
14	337.5	52	16	56.5	84	393.5	384	274
15	356	52	2	3	159	390.5	384	274
16	179	292.5	274	188.5	272	224.5	384	40
17	209.5	145.5	170	188.5	196	195	384	274
18	283	145.5	167.5	174	204	195	384	274
19	223.5	145.5	162.5	256	237	162.5	384	274
20	129	292.5	331	235	325.5	258	384	40
21	145.5	292.5	331	391.5	385.5	19	384	114.5
22	164	292.5	281	245.5	218	89.5	384	114.5
23	179	292.5	289	214.5	259	258	384	114.5
24	164	292.5	299	188.5	271	89.5	384	114.5
25	145.5	292.5	246	113	295	350.5	384	114.5
26	223.5	145.5	167.5	214.5	224	224.5	384	274
27	396.5	52	27.5	7	16	288.5	384	274
28	396.5	52	75	11	30	224.5	384	274

... Data | Covariance | Pearson | **Ranks** | Spearman | Sheet8

- Because horsepower is missing some values, they cannot be ranked and so appear as #N/A. Since there are only a few of them, we can remove them manually.
- This will avoid errors when calculating the Pearson coefficient in the next step, exactly as we did before; the result is as follows:

	A	B	C	D	E	F	G	H	I
1		*Rank_mpg*	*Rank_cylinders*	*Rank_displacement*	*Rank_horsepower*	*Rank_weight*	*Rank_acceleration*	*Rank_model_year*	*Rank_origin*
2	Rank_mpg	1							
3	Rank_cylinders	-0.821864491	1						
4	Rank_displacement	-0.855692012	0.911875915	1					
5	Rank_horsepower	-0.853320216	0.815689638	0.875770352	1				
6	Rank_weight	-0.874947398	0.873313559	0.945985564	0.878284909	1			
7	Rank_acceleration	0.43867748	-0.474189066	-0.496511921	-0.657631236	-0.404550372	1		
8	Rank_model_year	0.573468703	-0.335012387	-0.30525727	-0.389975332	-0.277014582	0.274632098	1	
9	Rank_origin	0.580693694	-0.604550452	-0.707196539	-0.509090776	-0.628434003	0.220573847	0.166551172	1
10									

Studying the Spearman's correlation