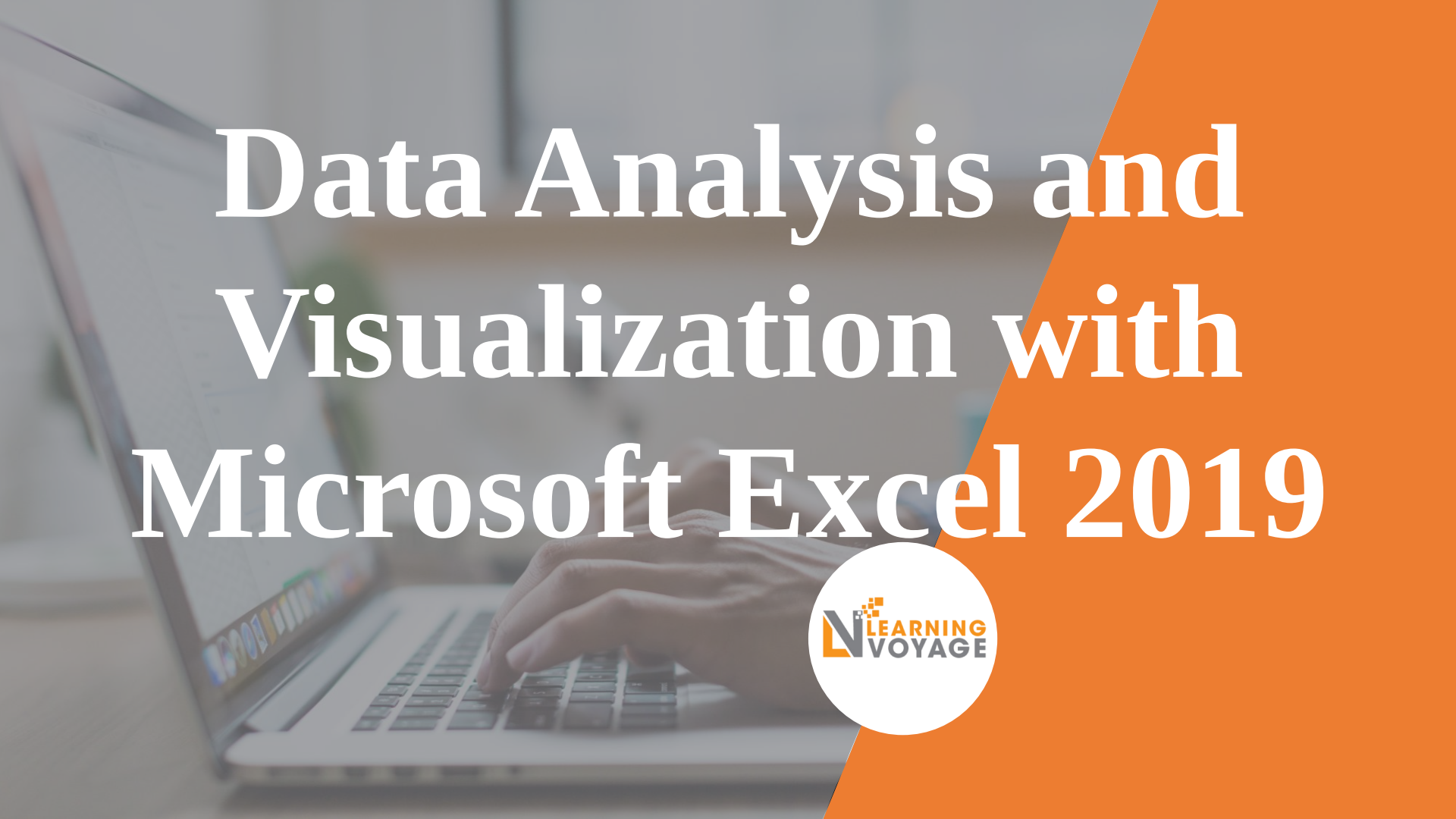
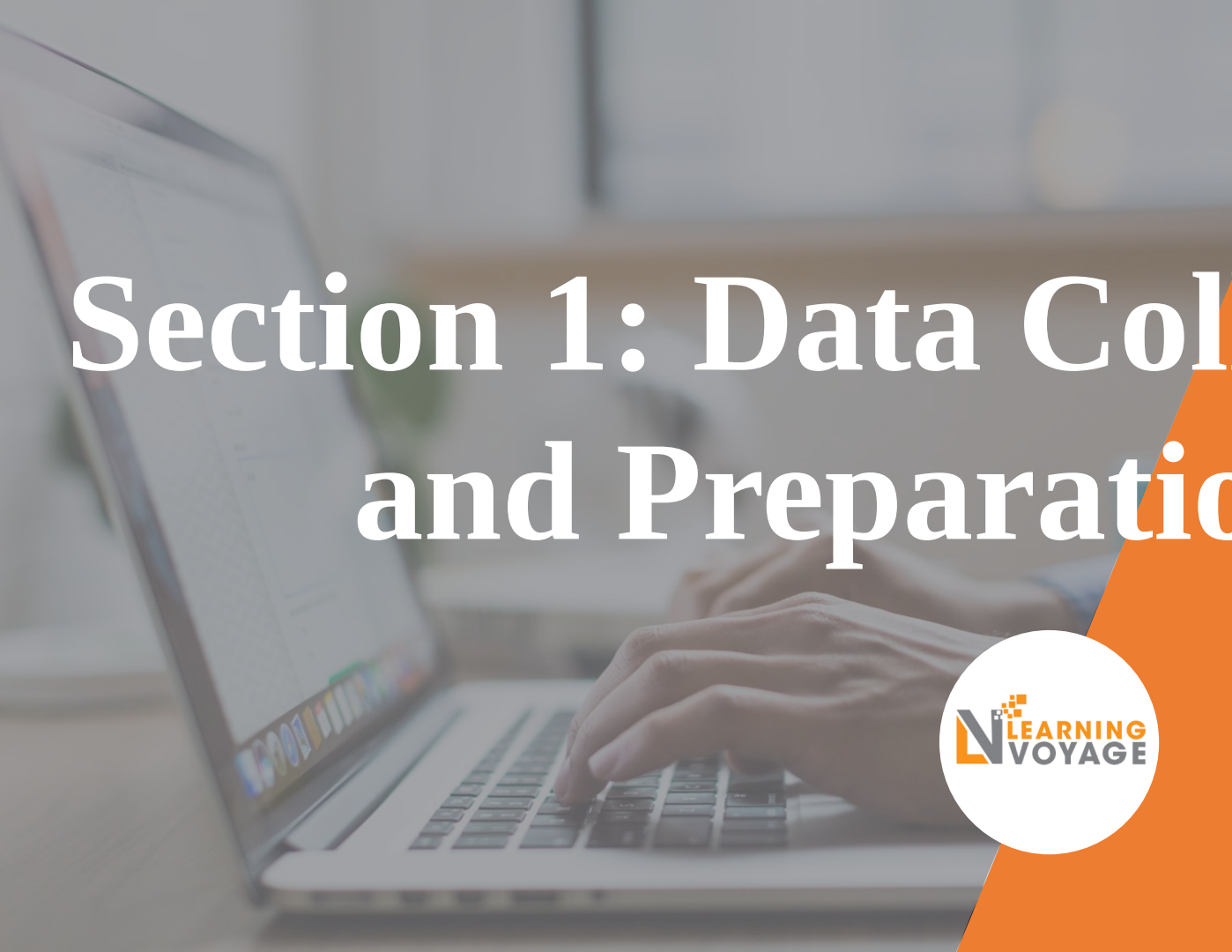


Data Analysis and Visualization with Microsoft Excel 2019

LEARNING VOYAGE



Section 1: Data Collection and Preparation

LEARNING VOYAGE

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

2: 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:

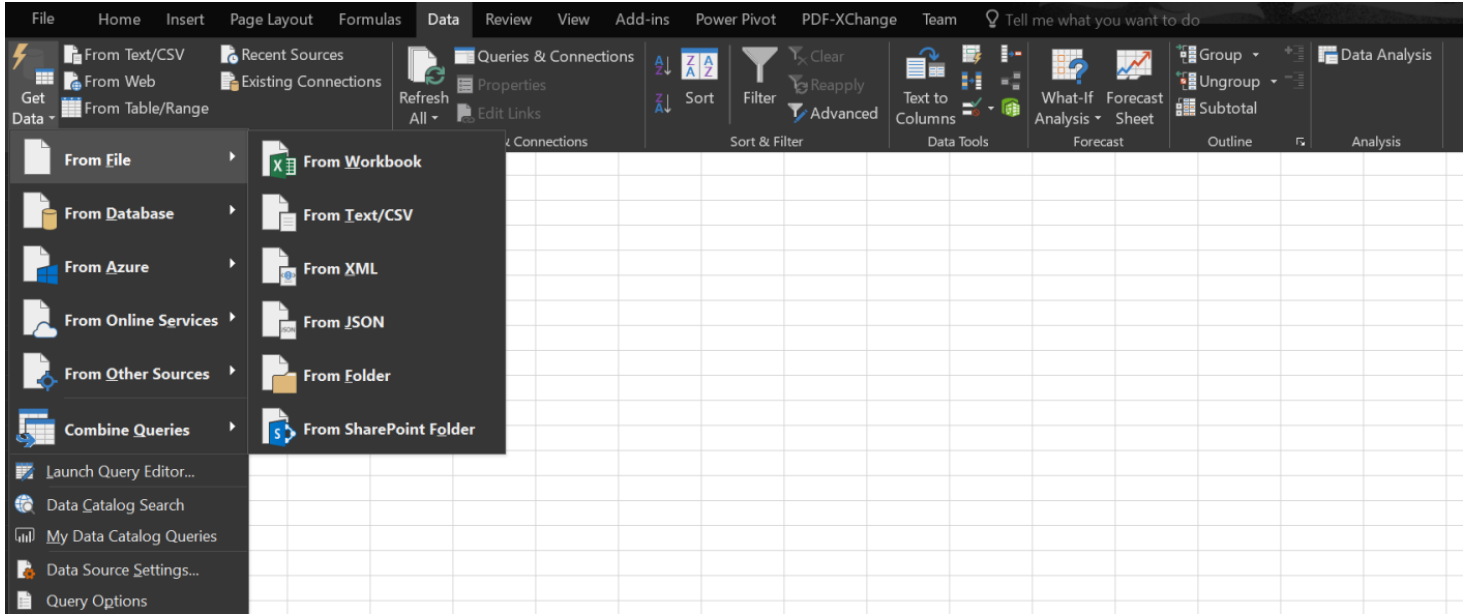
- Importing data from a text file
- Importing data from another Excel workbook
- Importing data from a web page
- Importing data from Facebook
- Importing data from a JSON file
- Importing data from a database

Technical requirements

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

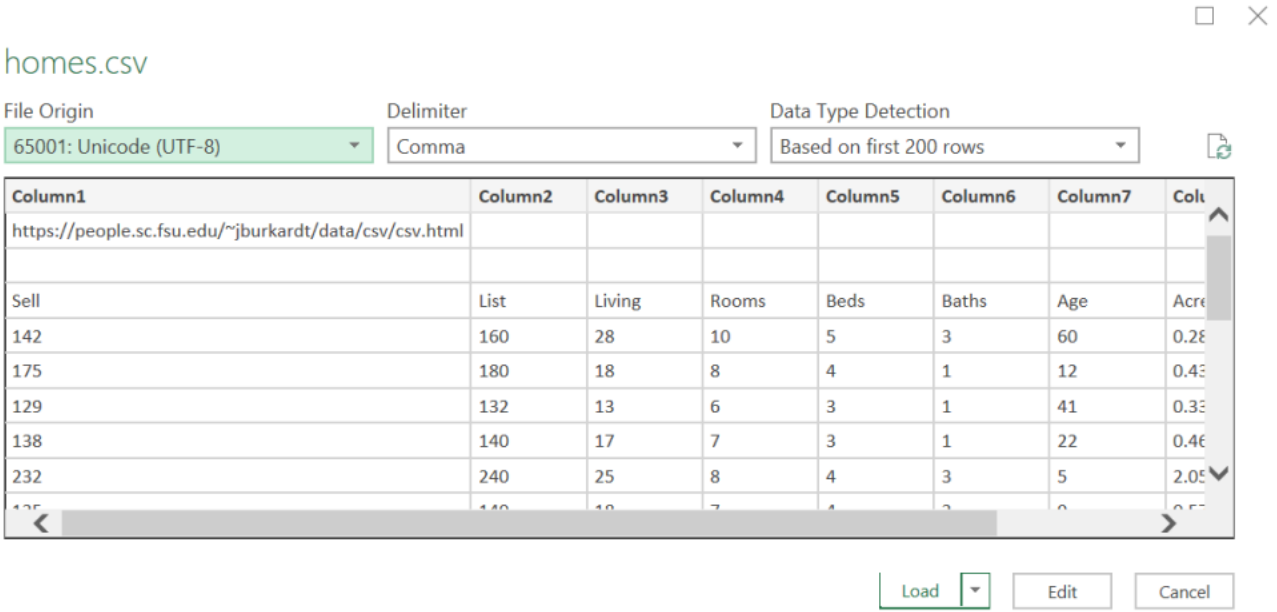
Importing data from a text file

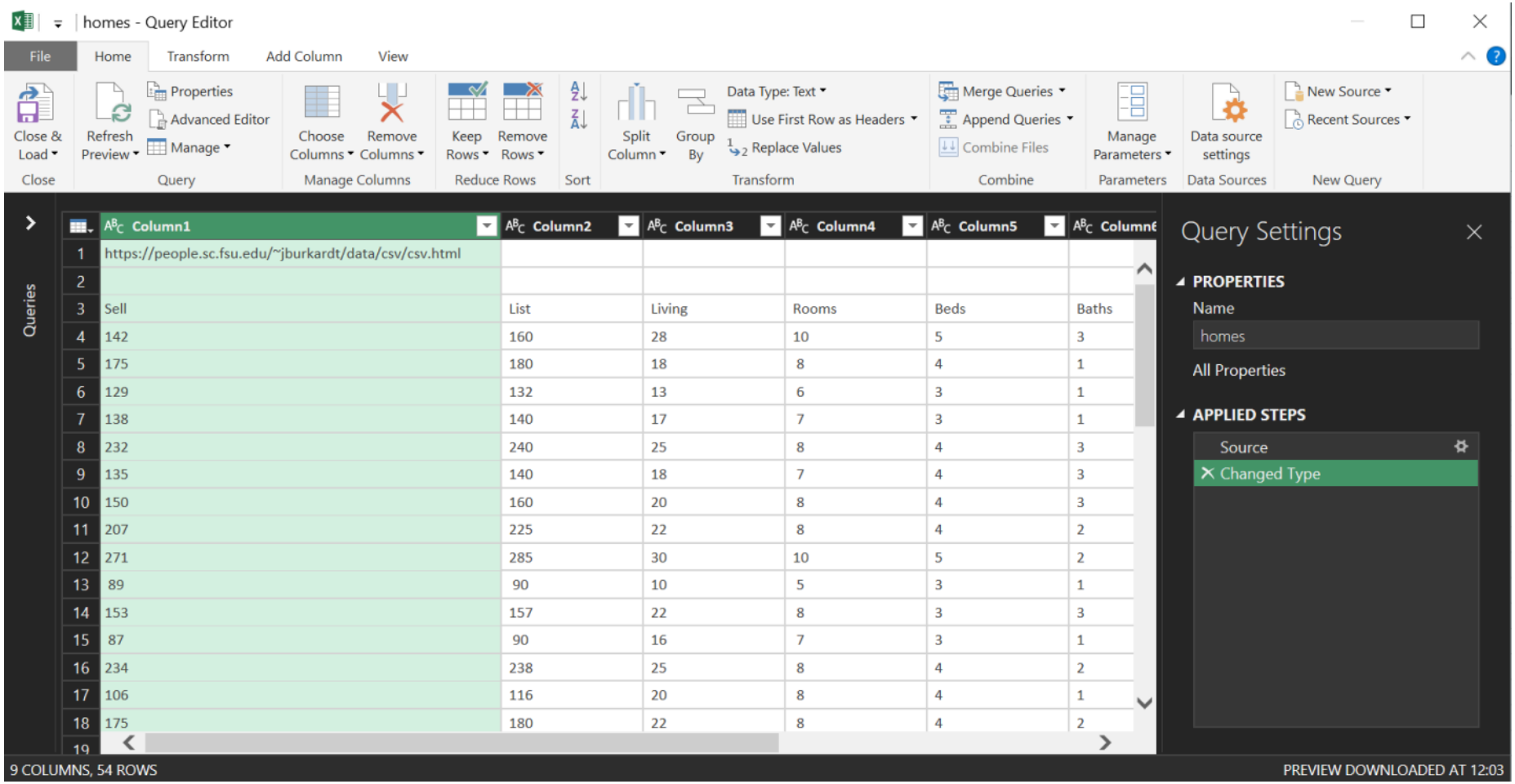
- Click on Data.
- Navigate to Get Data | From File | From Text/CSV:



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:

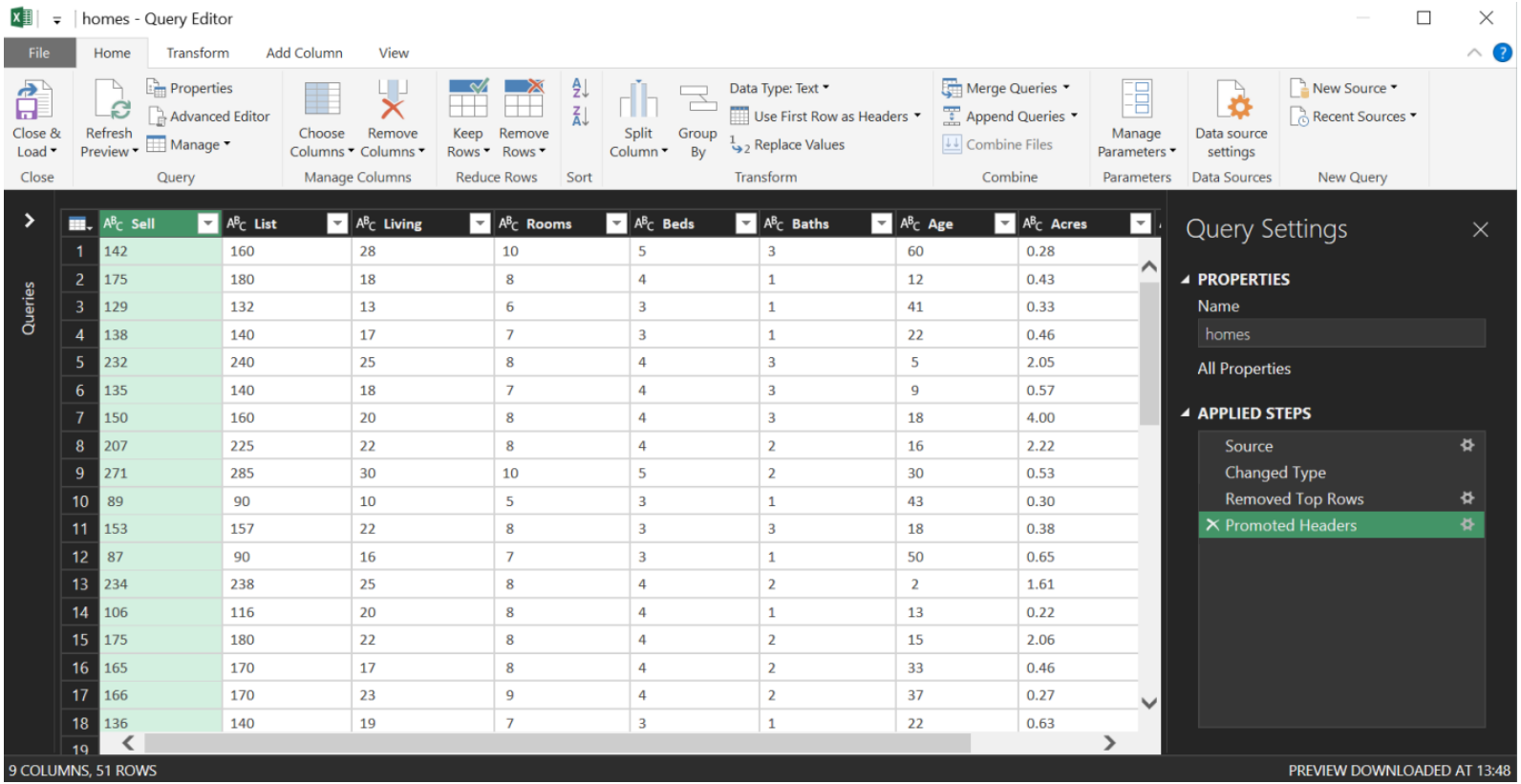



homes - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Transform
Merge Queries
Append Queries
Combine Files
Combine
Manage Parameters
Parameters
Data source settings
Data Sources
New Source
Recent Sources
New Query
Queries
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:

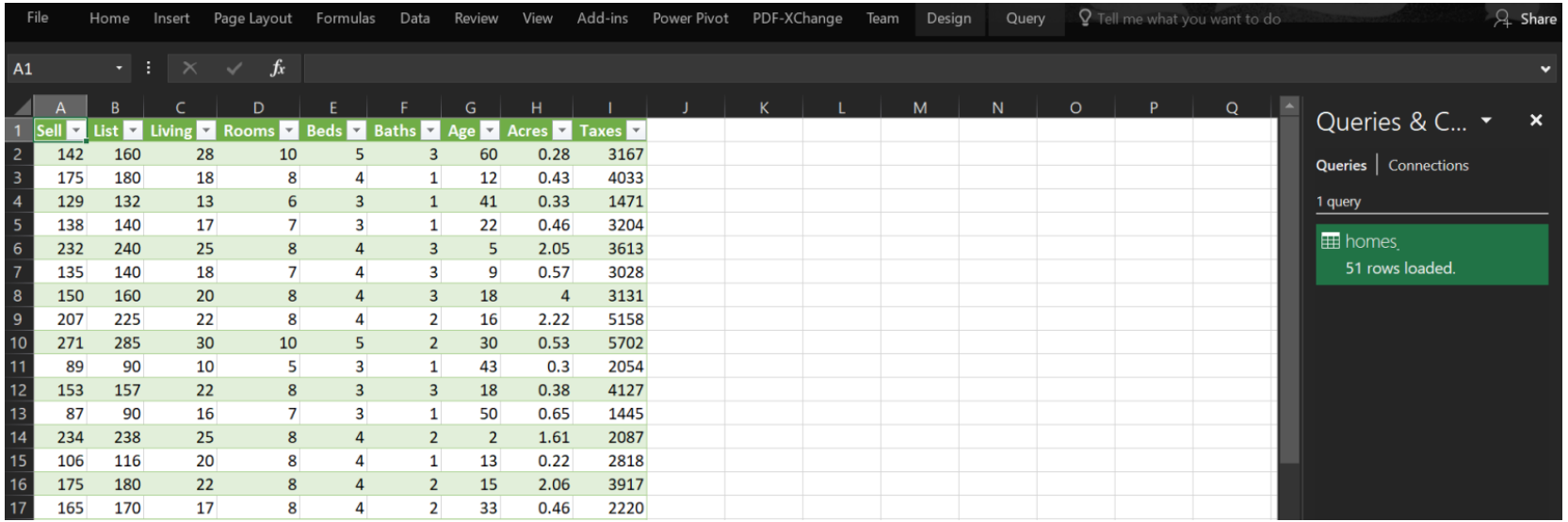


Importing data from a text file

- 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:



	A	B	C	D	E	F	G	H	I
1	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

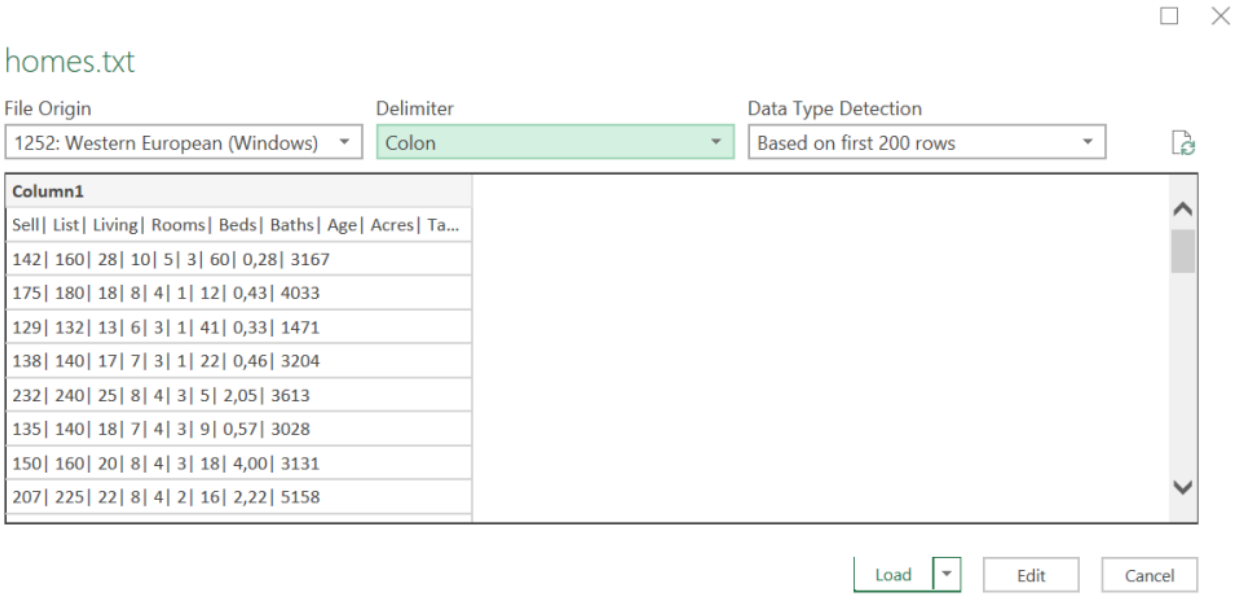
Importing data from a text file

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

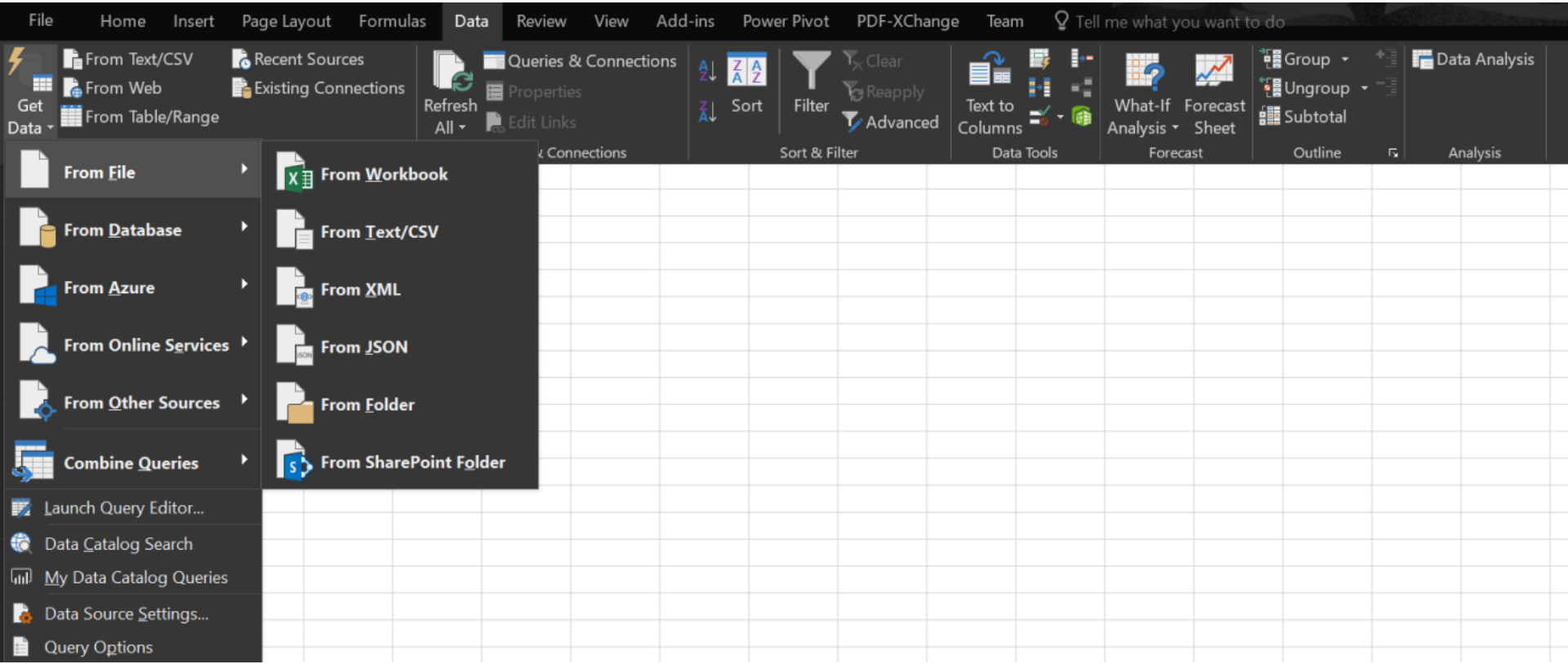
Sell	List	Living	Rooms	Beds	Baths	Age	Acres	Taxes
142	160	28	10	5	3	60	0.28	3167
175	180	18	8	4	1	12	0.43	4033
129	132	13	6	3	1	41	0.33	1471
138	140	17	7	3	1	22	0.46	3204
232	240	25	8	4	3	5	2.05	3613
135	140	18	7	4	3	9	0.57	3028
150	160	20	8	4	3	18	4	3131
207	225	22	8	4	2	16	2.22	5158

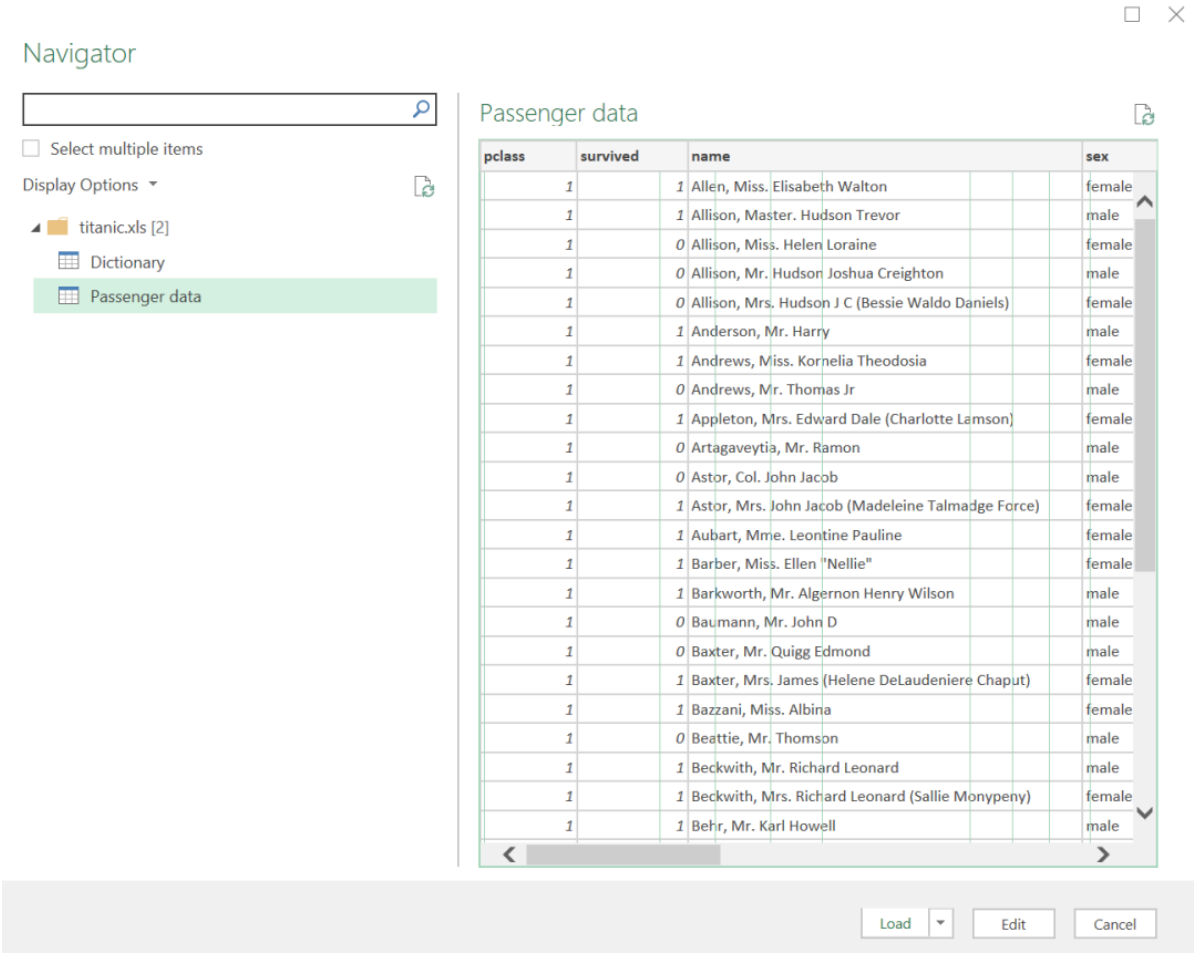
Load | Edit | Cancel

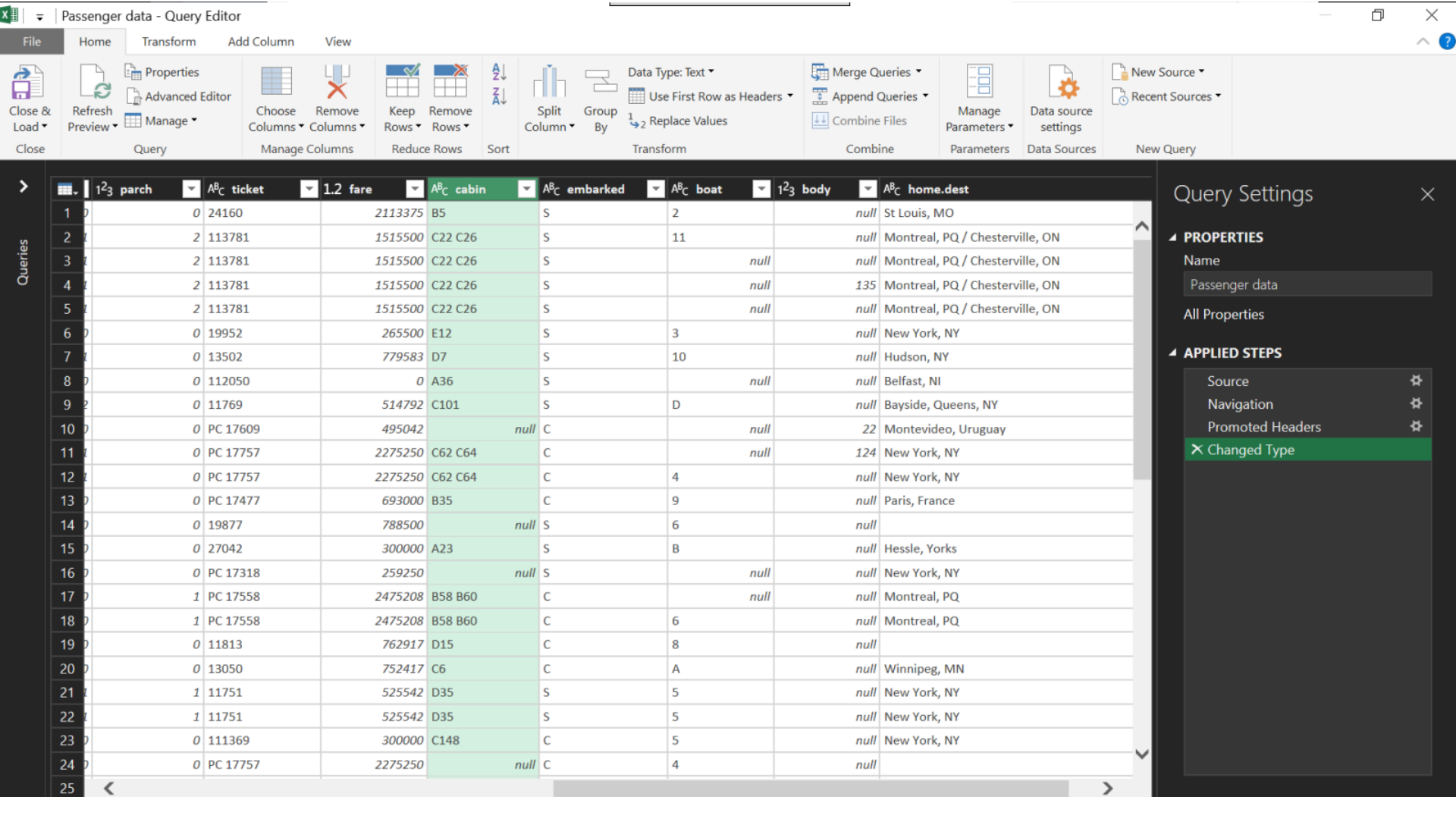
Importing data from another Excel workbook

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

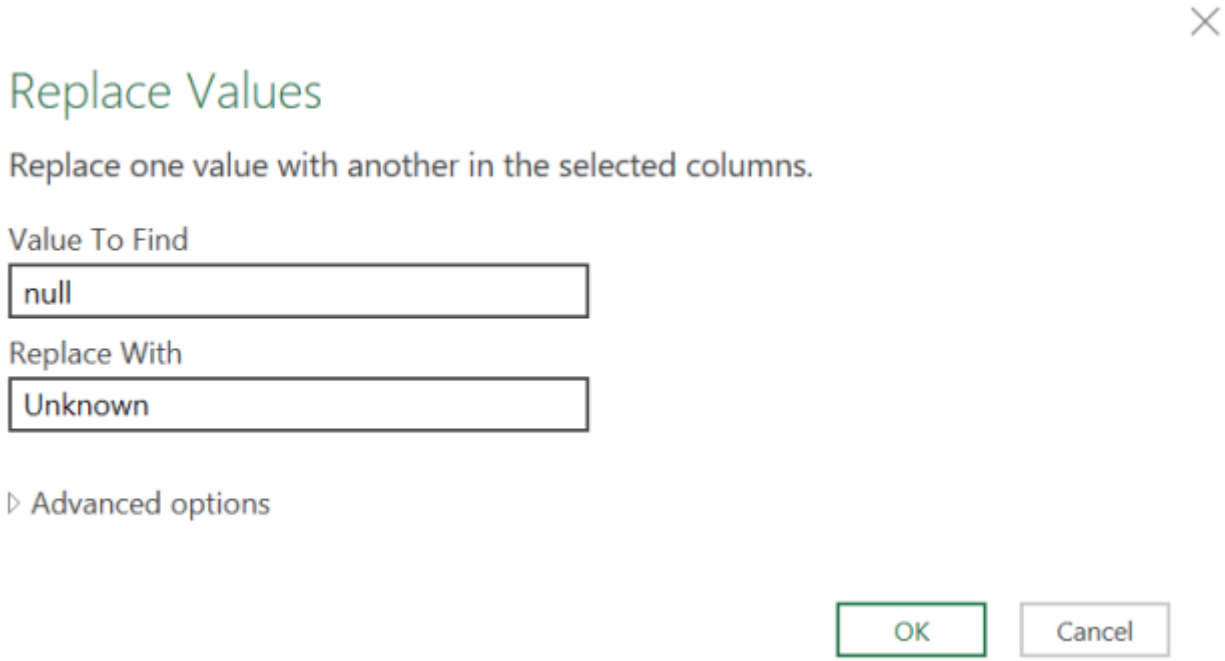
- Click on Data.
- Navigate to Get Data | From File | From Workbook, as shown in the following screenshot on next slide.


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 0 24160 2113375 B5 S 2 null St Louis, MO
2 1 2 113781 1515500 C22 C26 S 11 null Montreal, PQ / Chesterville, ON
3 1 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
4 1 2 113781 1515500 C22 C26 S null 135 Montreal, PQ / Chesterville, ON
5 1 2 113781 1515500 C22 C26 S null null Montreal, PQ / Chesterville, ON
6 0 0 19952 265500 E12 S 3 null New York, NY
7 1 0 13502 779583 D7 S 10 null Hudson, NY
8 0 0 112050 0 A36 S null null Belfast, NI
9 2 0 11769 514792 C101 S D null Bayside, Queens, NY
10 0 0 PC 17609 495042 null C null 22 Montevideo, Uruguay
11 1 0 PC 17757 2275250 C62 C64 C null 124 New York, NY
12 1 0 PC 17757 2275250 C62 C64 C 4 null New York, NY
13 0 0 PC 17477 693000 B35 C 9 null Paris, France
14 0 0 19877 788500 null S 6 null
15 0 0 27042 300000 A23 S B null Hessle, Yorks
16 0 0 PC 17318 259250 null S null null New York, NY
17 0 1 PC 17558 2475208 B58 B60 C null null Montreal, PQ
18 0 1 PC 17558 2475208 B58 B60 C 6 null Montreal, PQ
19 0 0 11813 762917 D15 C 8 null
20 0 0 13050 752417 C6 C A null Winnipeg, MN
21 1 1 11751 525542 D35 S 5 null New York, NY
22 1 1 11751 525542 D35 S 5 null New York, NY
23 0 0 111369 300000 C148 C 5 null New York, NY
24 0 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

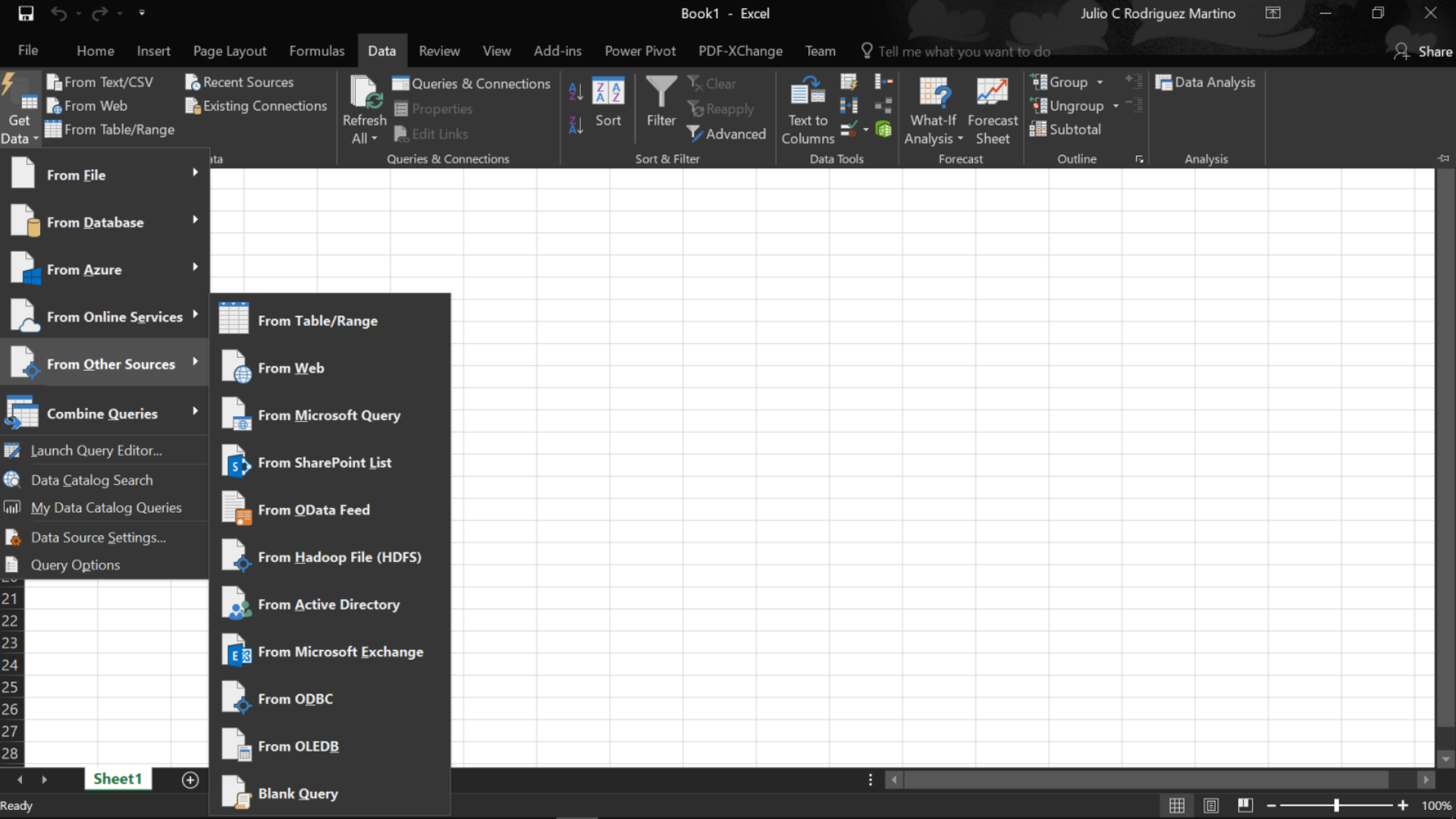
All Properties

APPLIED STEPS

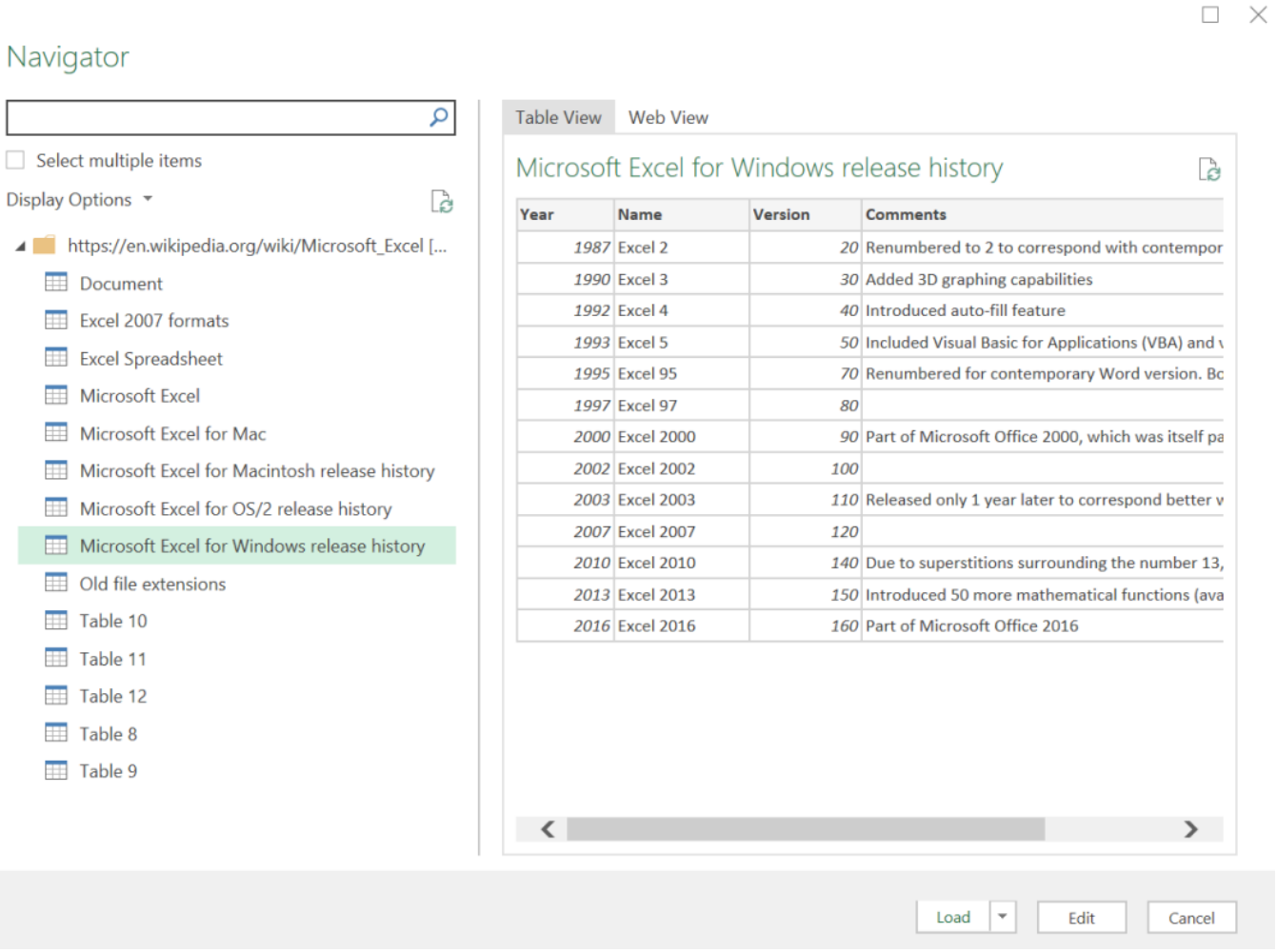
- Source
- Navigation
- Promoted Headers
- Changed Type
- Replaced Value

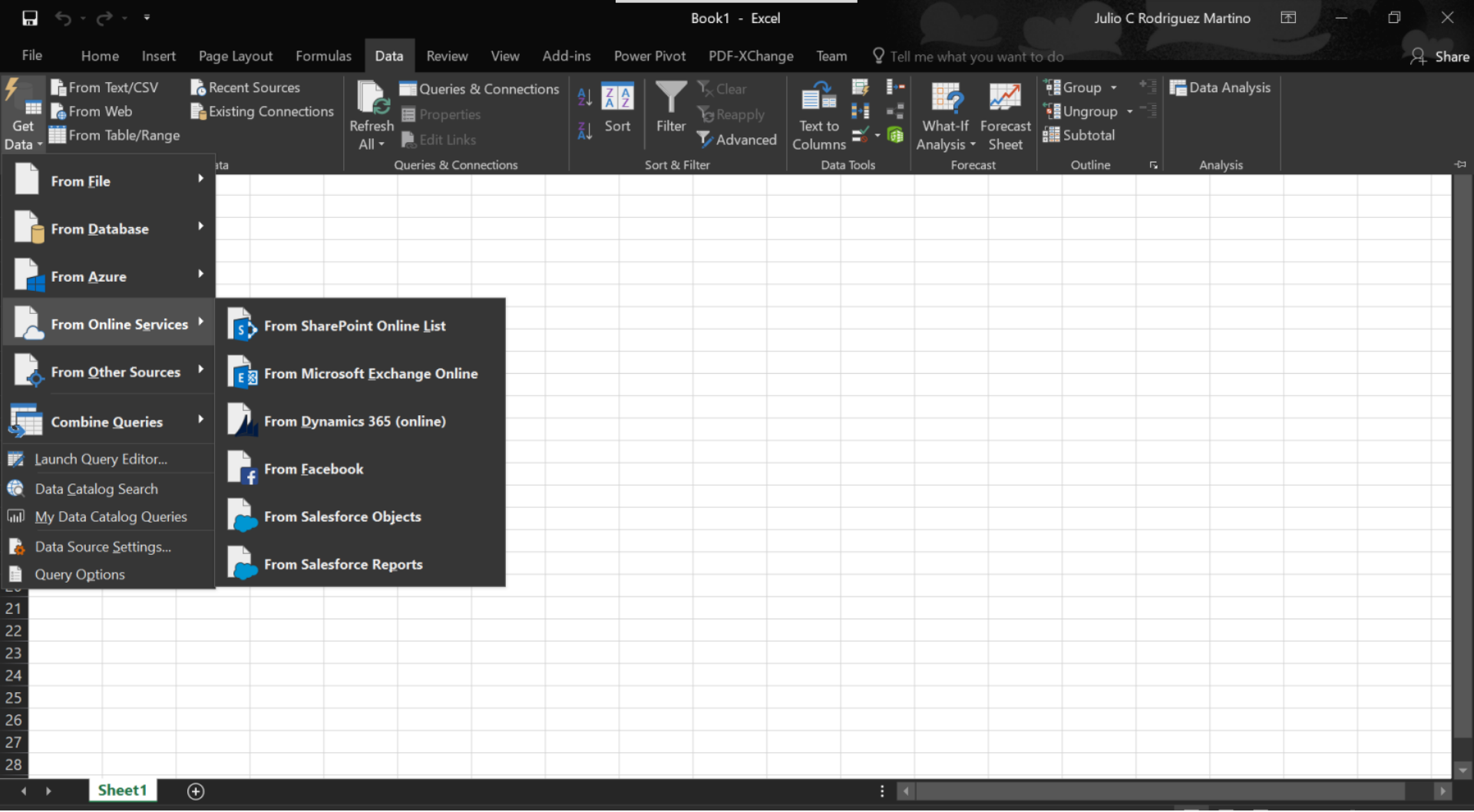
14 COLUMNS, 999+ ROWS

PREVIEW DOWNLOADED AT 07:42

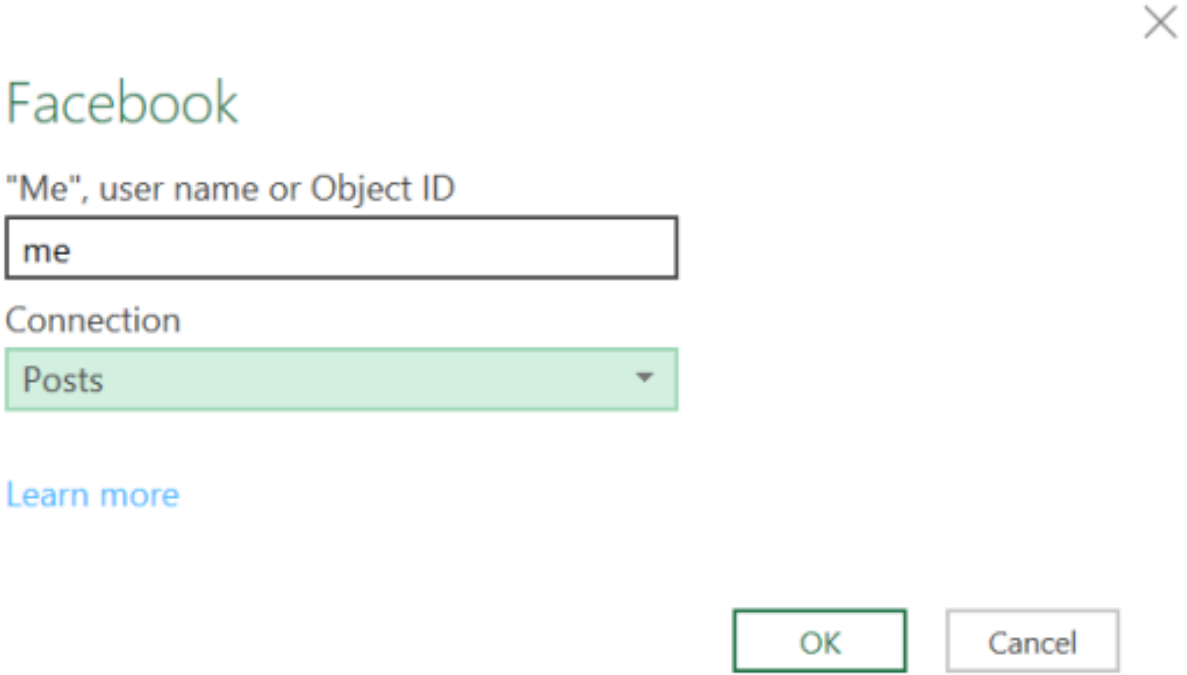

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
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 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%

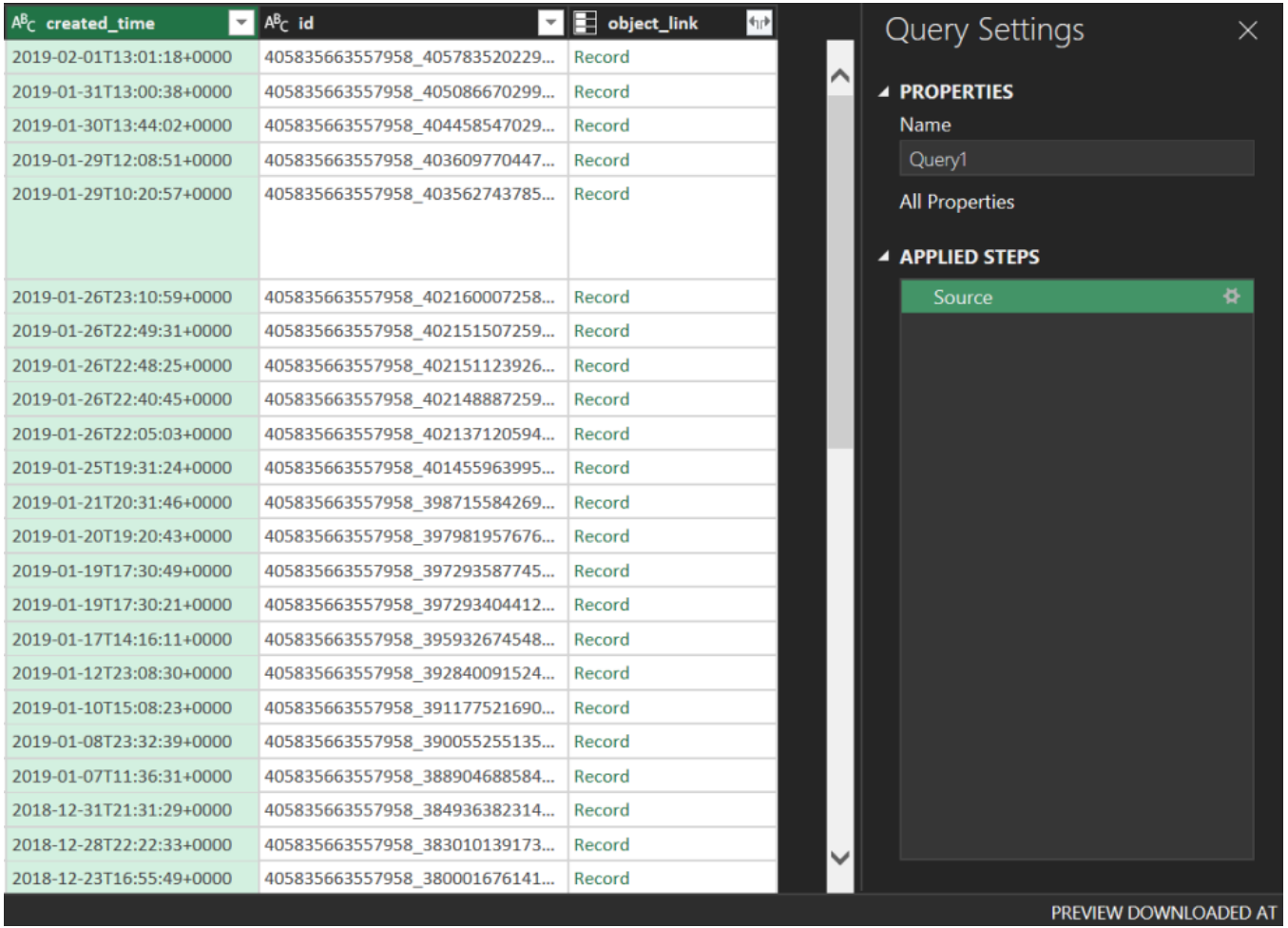
Importing data from a web page


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

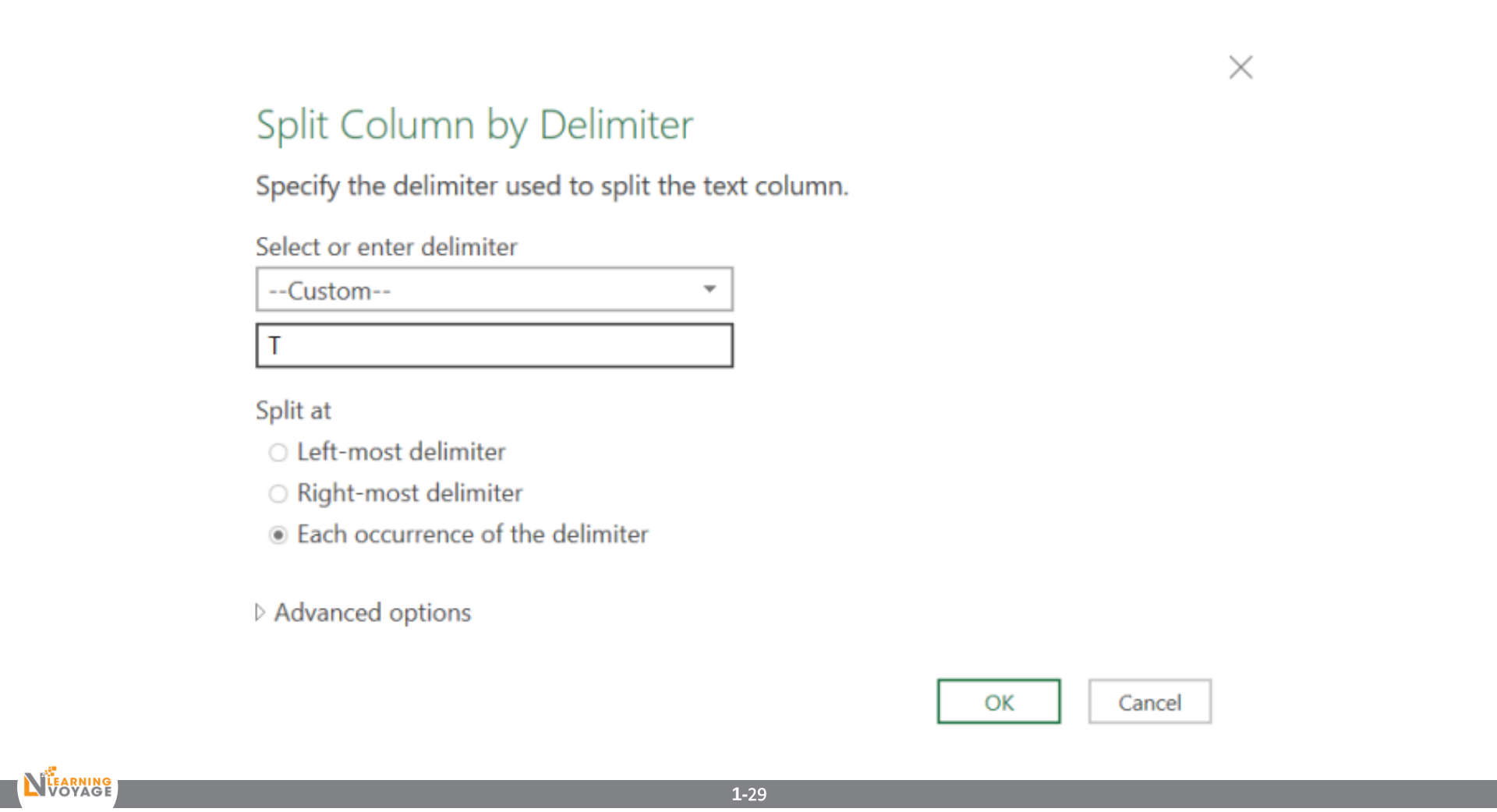

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit 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

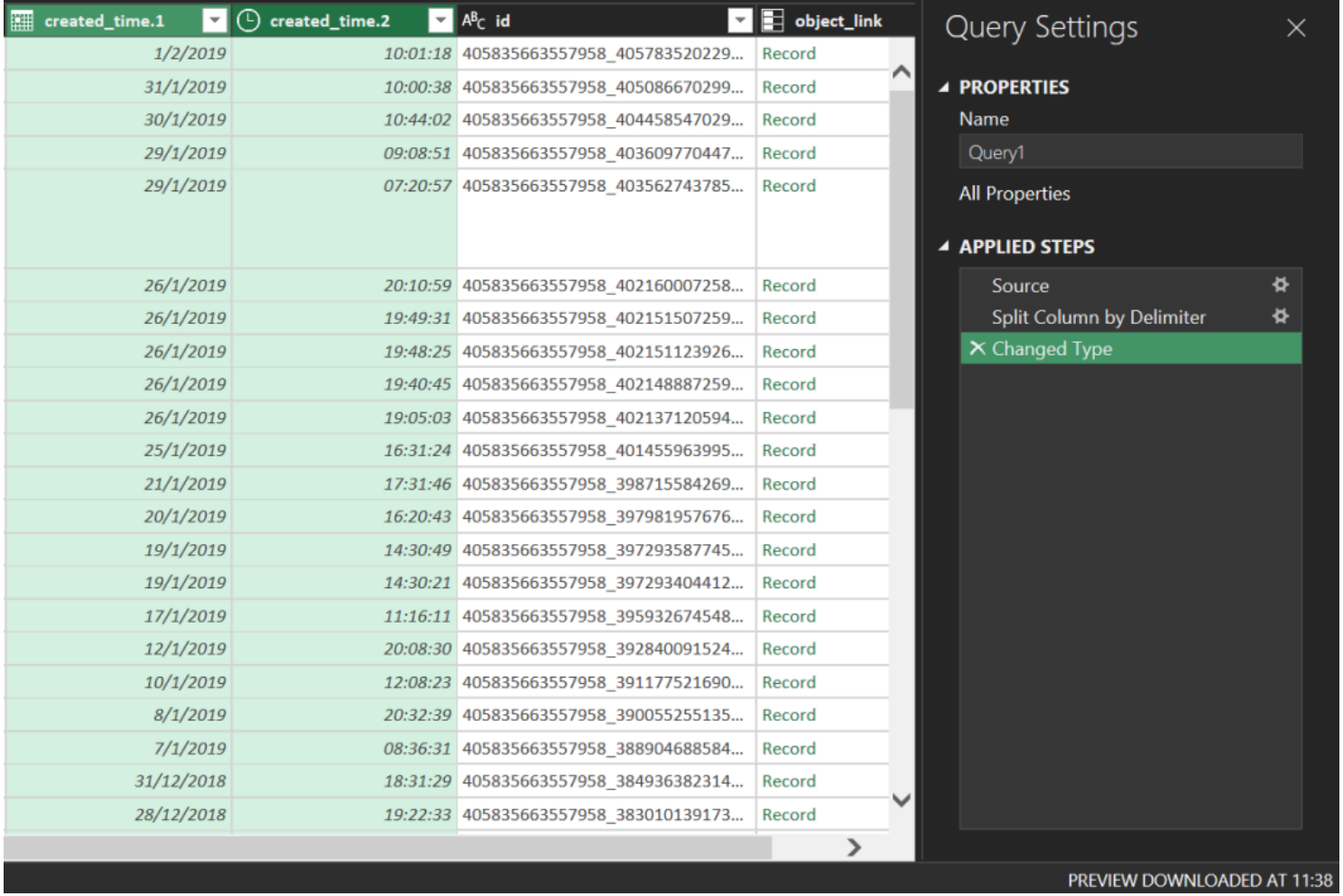
Importing data from Facebook





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!

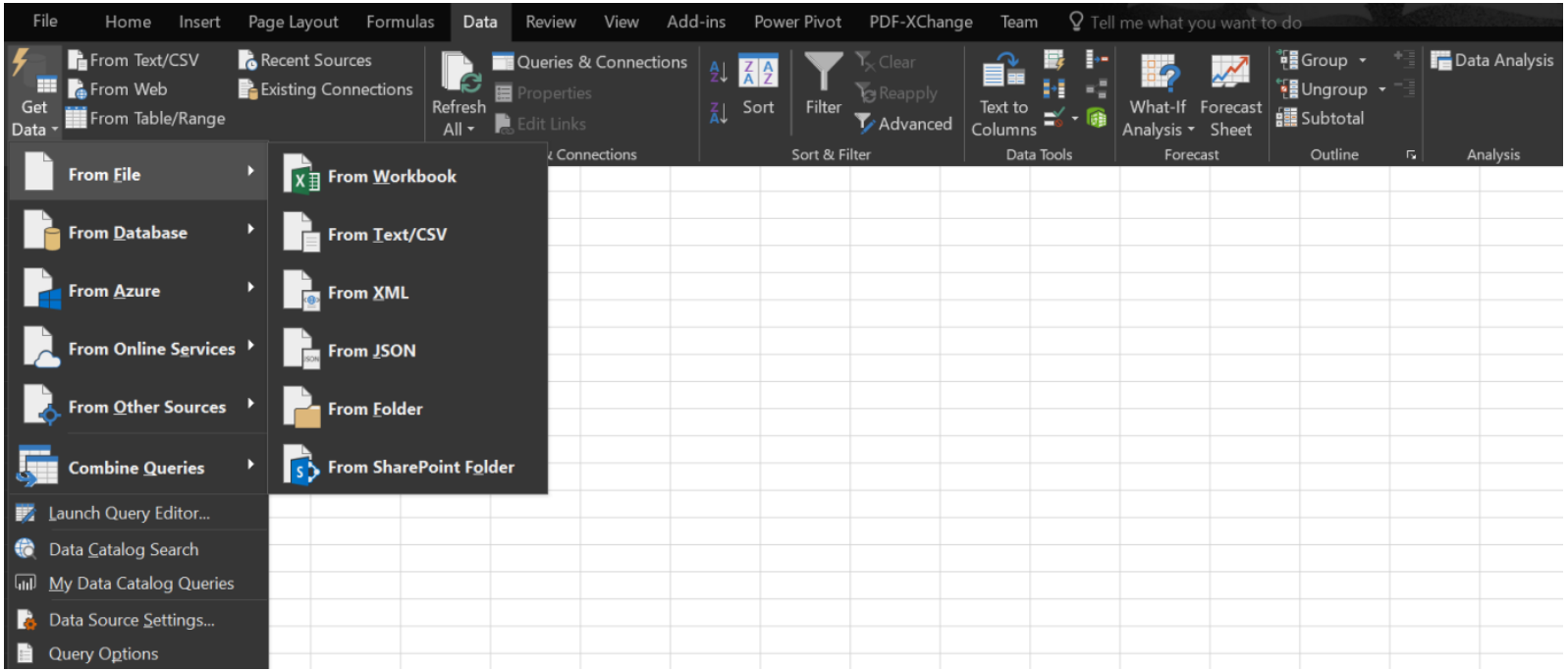
Analyze

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

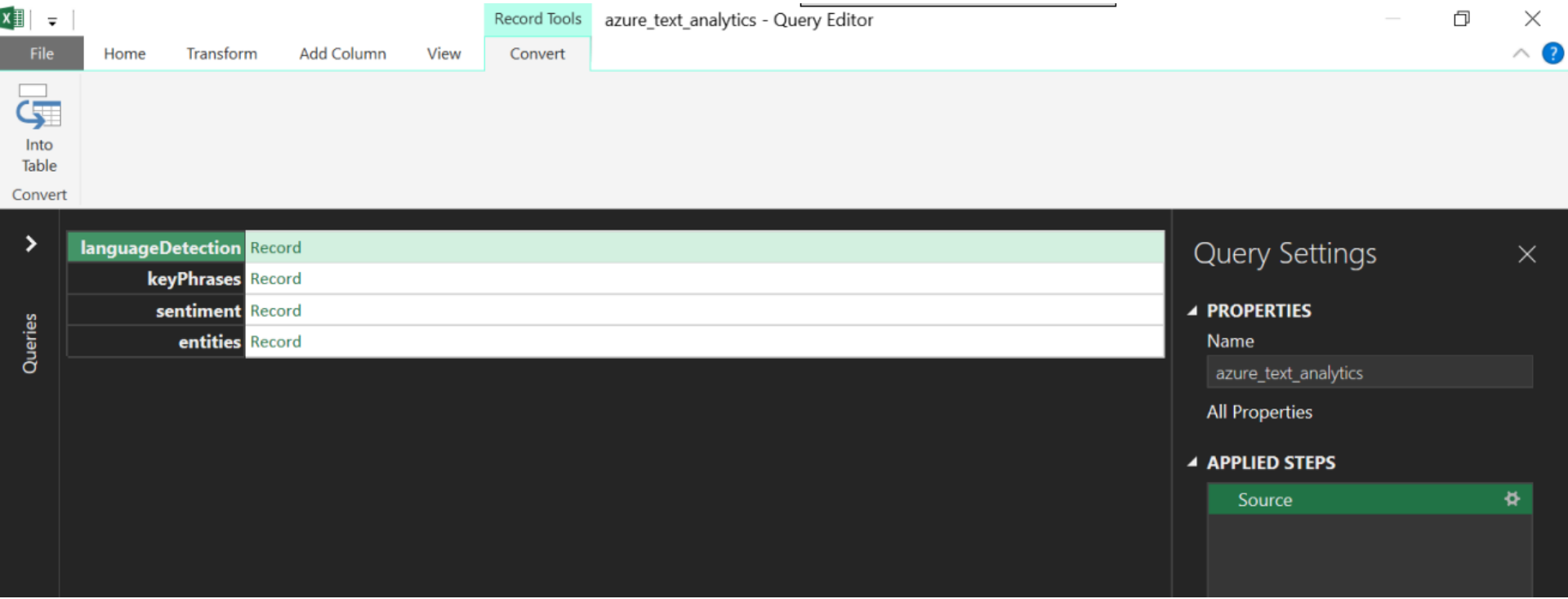
Analyzed text | JSON


```
{
  "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:

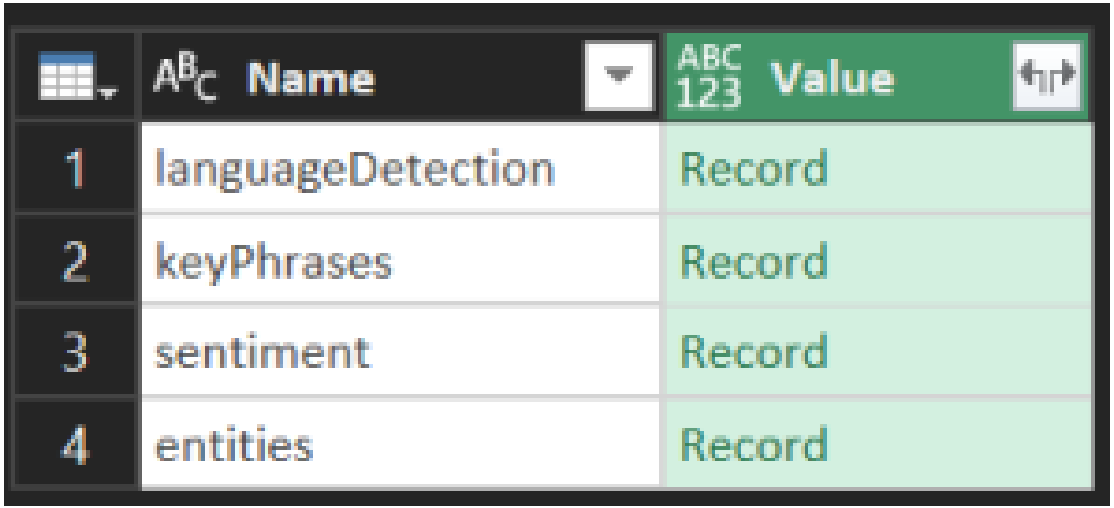


- You will get a preview showing the main fields in the JSON structure, as shown in the following screenshot:

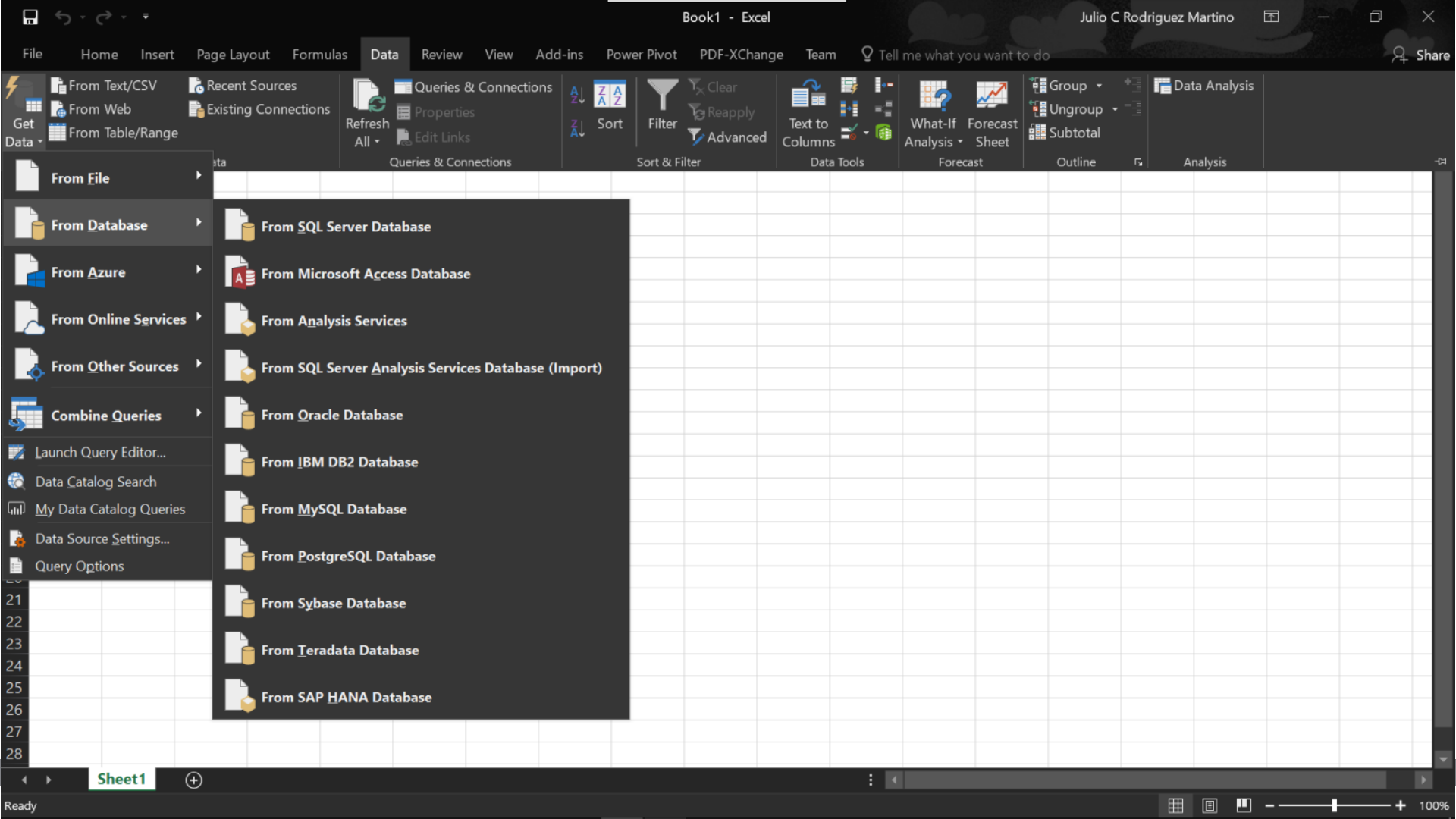


Importing data from a JSON file

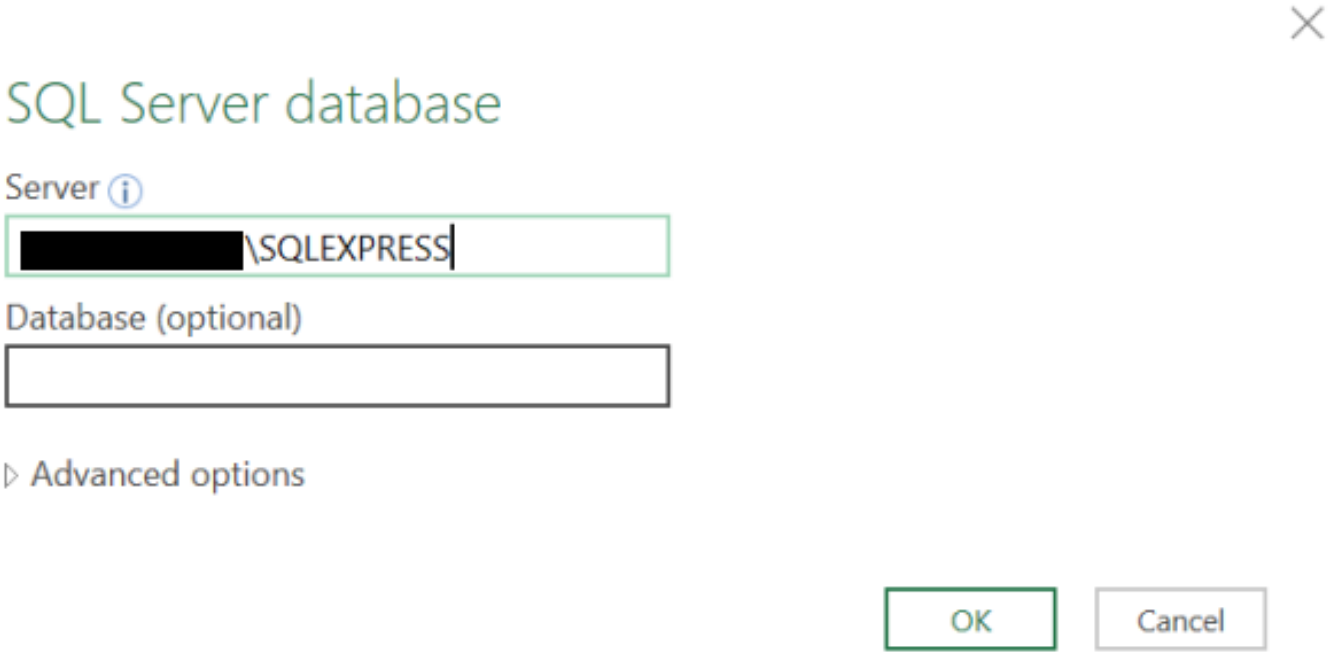
- Click on Into Table to convert the entries into regular Excel tables, as shown in the following screenshot



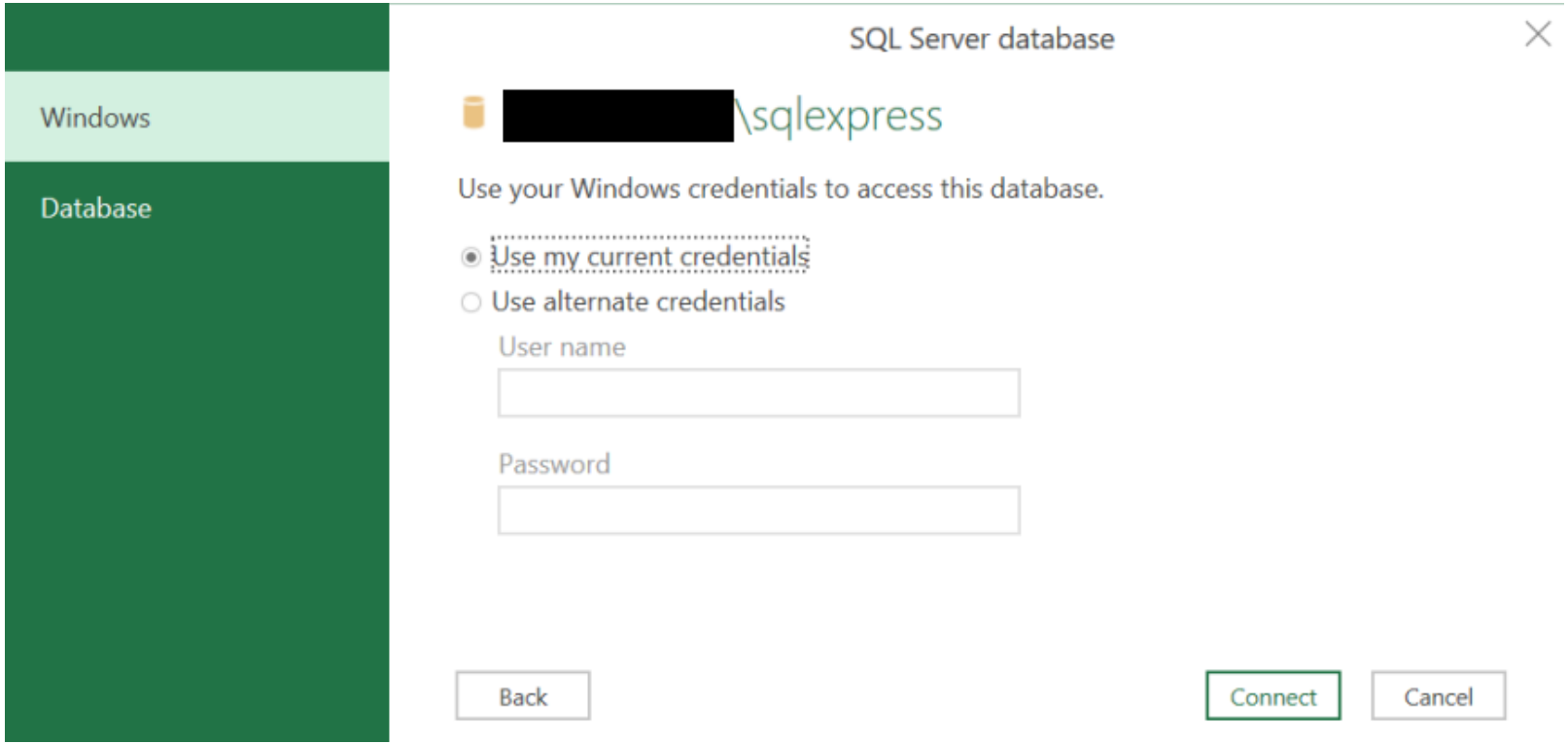
	Name	Value
1	languageDetection	Record
2	keyPhrases	Record
3	sentiment	Record
4	entities	Record

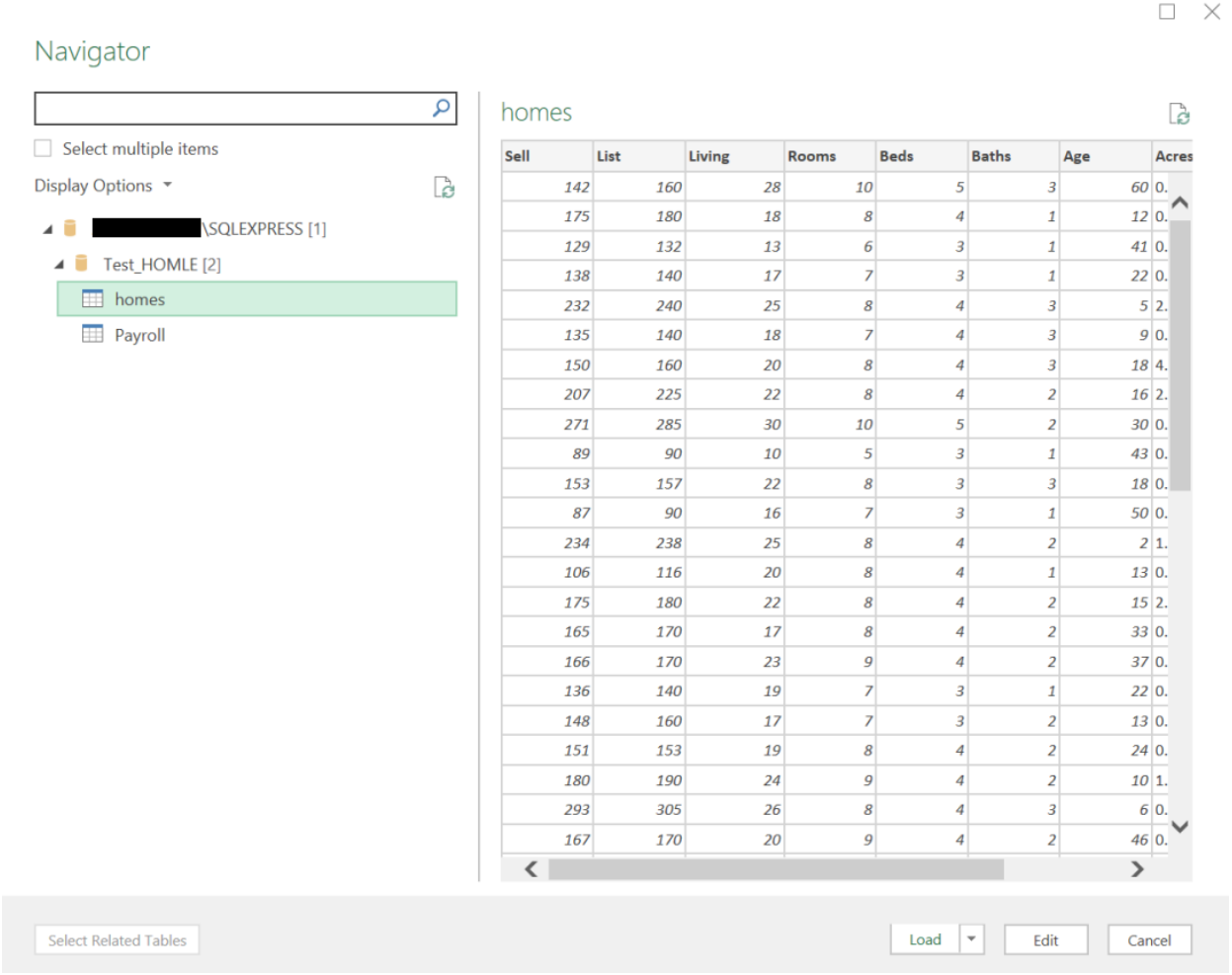

Book1 - Excel
Julio C Rodriguez Martino
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Share
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit 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



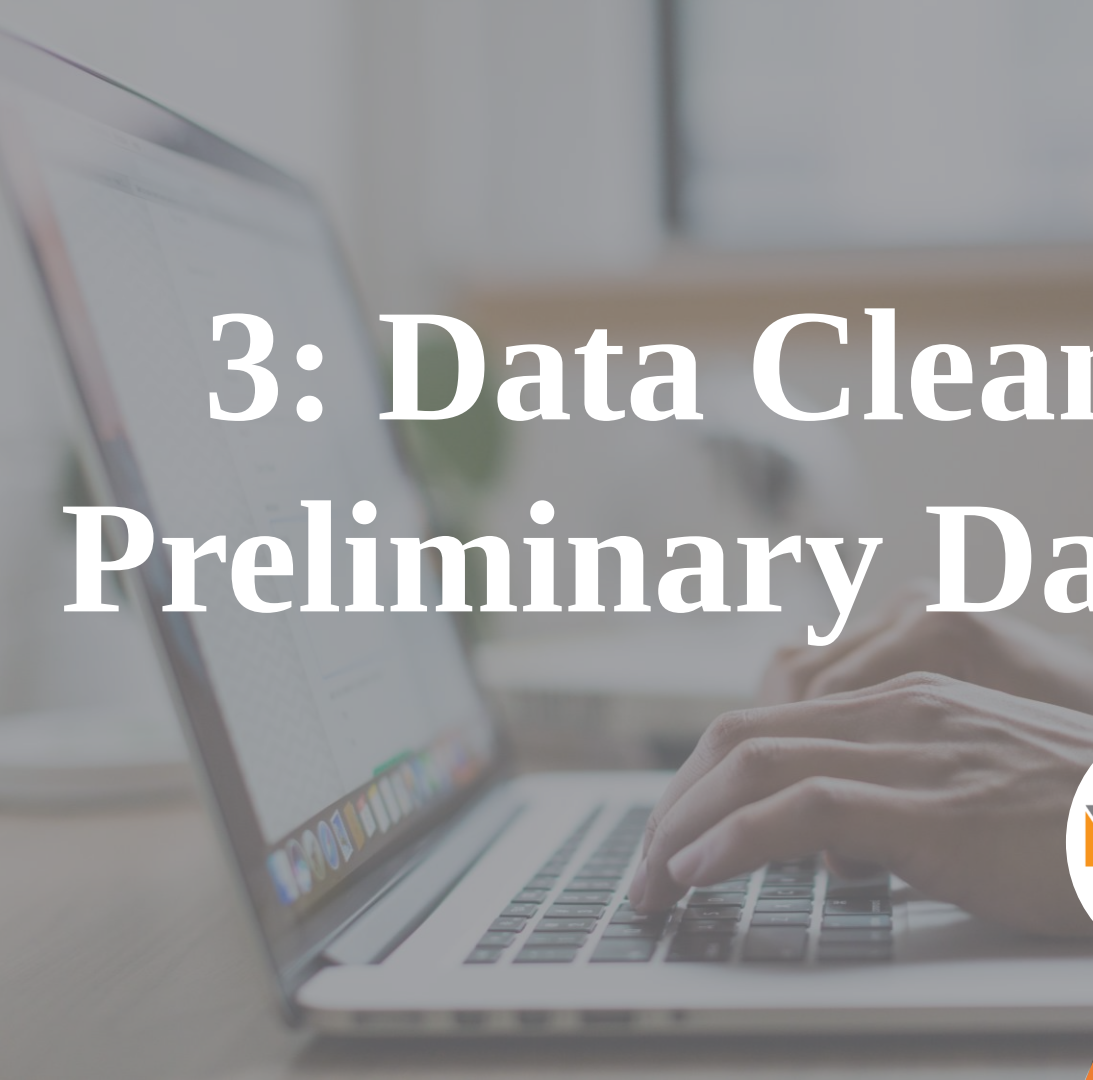
Importing data from a database




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.



3: Data Cleansing and Preliminary Data Analysis

LEARNING VOYAGE

Data Cleansing and Preliminary Data Analysis

In this lesson, we will cover the following topics:

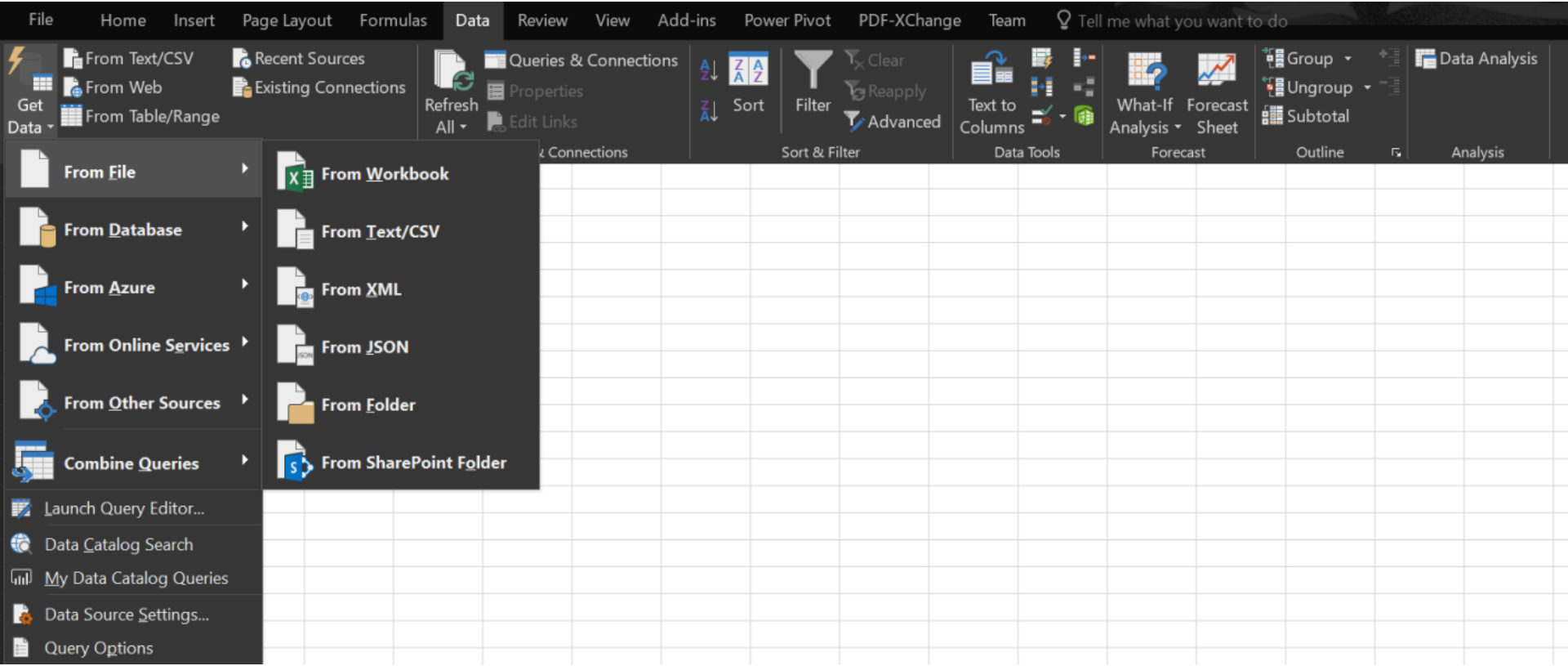
- Cleansing data
- Visualizing data for preliminary analysis
- Understanding unbalanced datasets

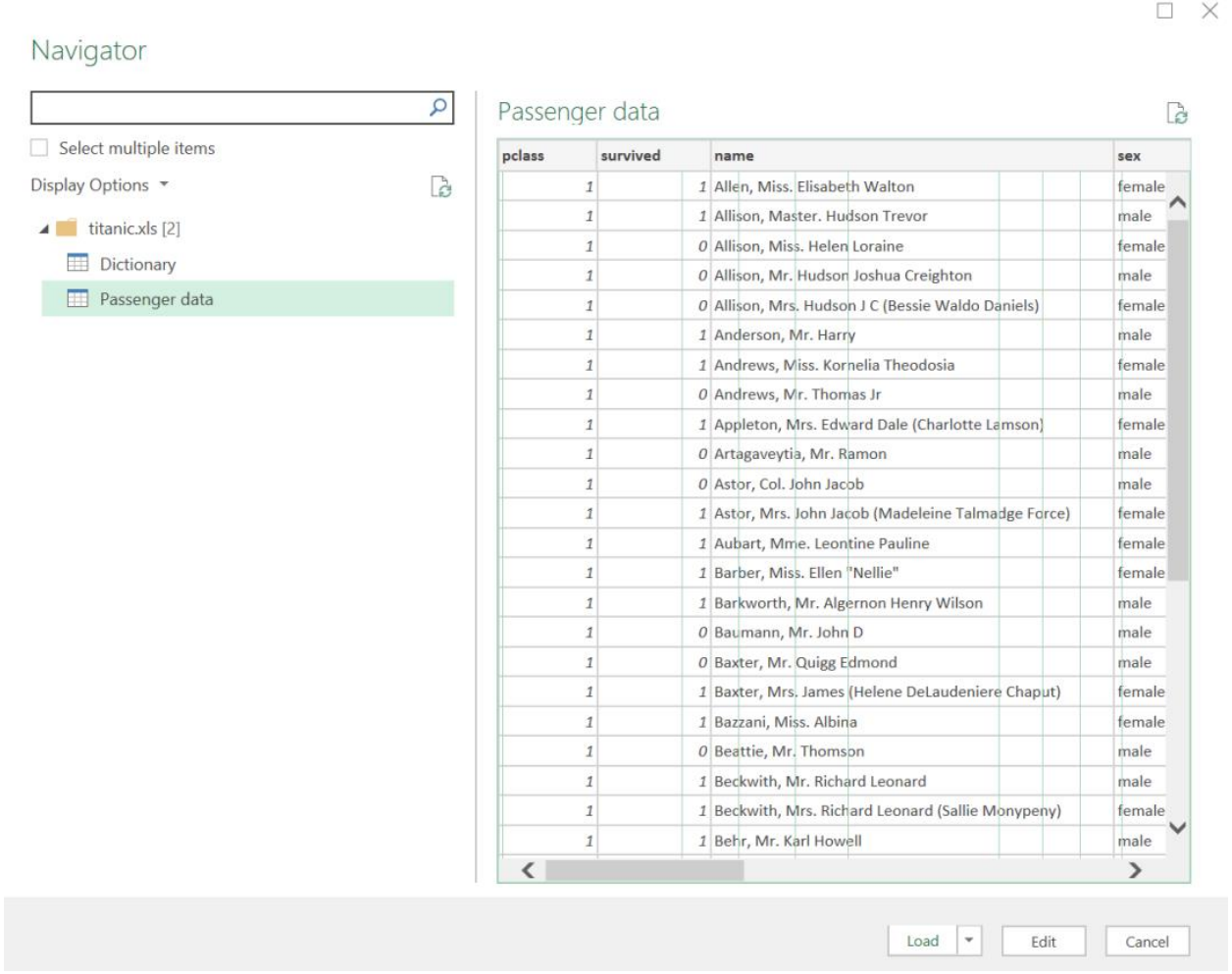
Technical requirements

- You will need to download the titanic.xls file

Cleansing data

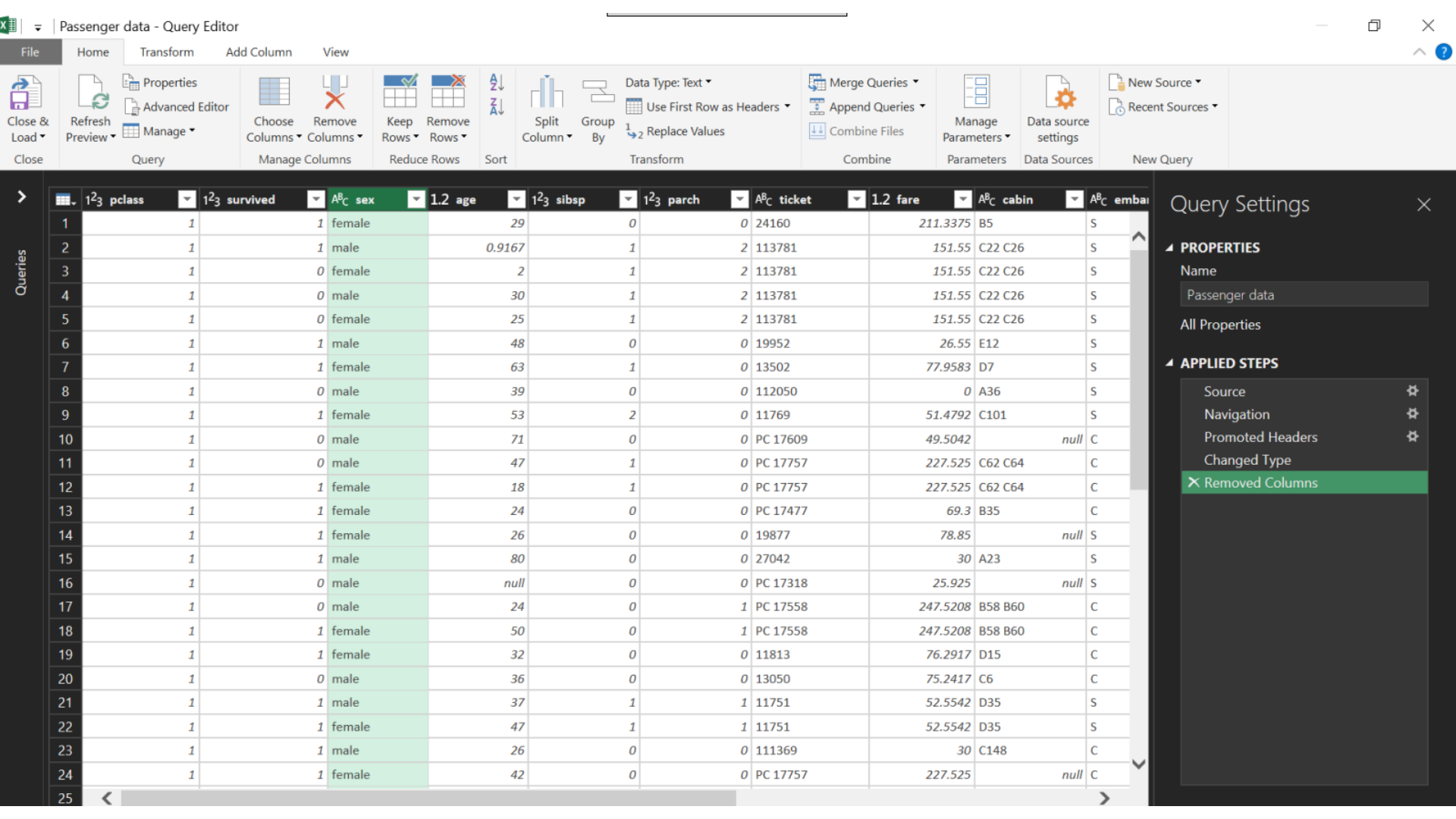
- Data is never clean – it always contains missing values, errors, incorrect formats, and other problems that make it impossible to feed to a machine learning model without preprocessing.
- This is what data cleansing is all about – correcting all these problems before starting the real analysis.
- As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

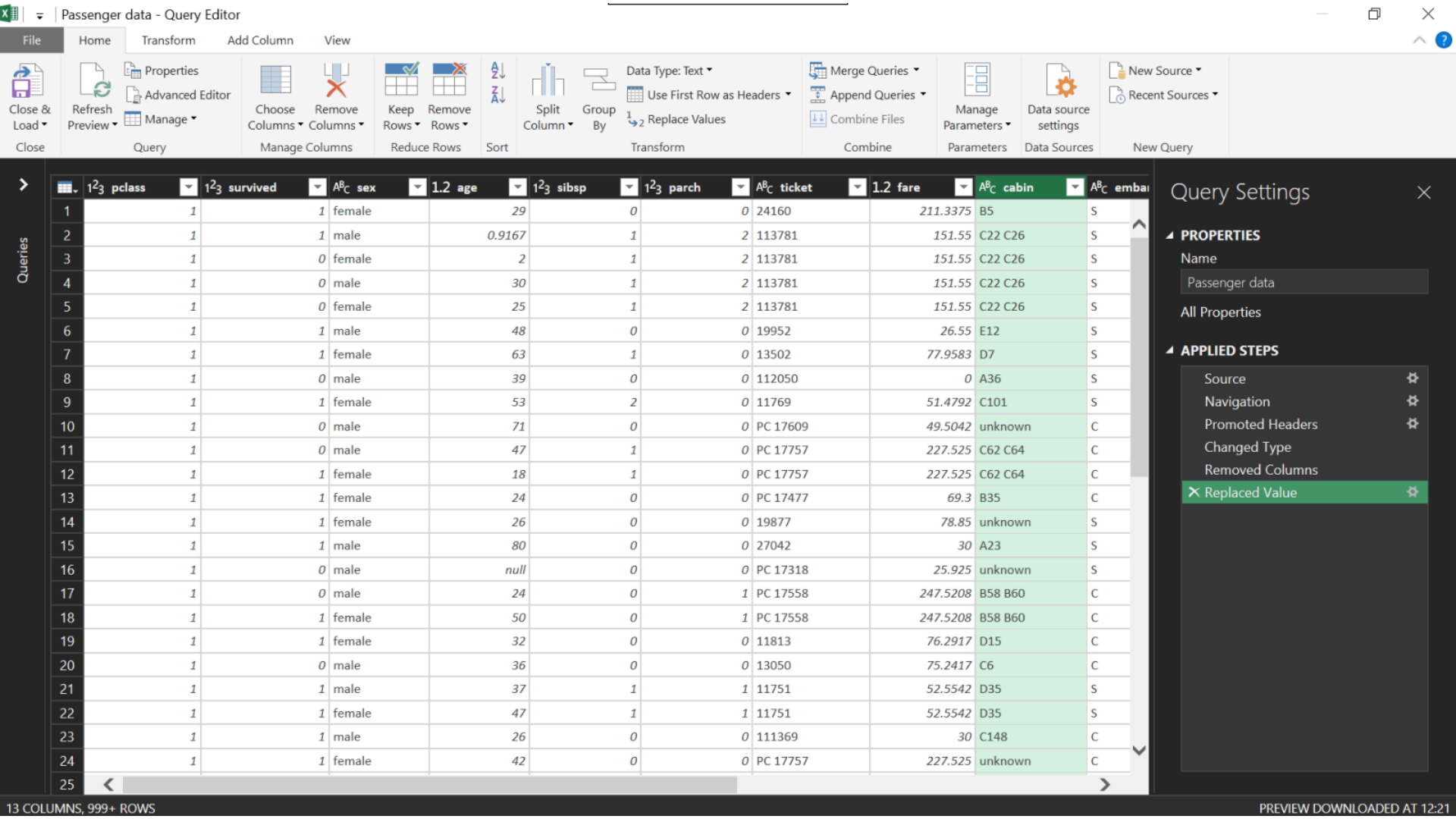

File
Home
Insert
Page Layout
Formulas
Data
Review
View
Add-ins
Power Pivot
PDF-XChange
Team
Tell me what you want to do
Get Data
From Text/CSV
From Web
From Table/Range
Recent Sources
Existing Connections
Refresh All
Queries & Connections
Properties
Edit Links
Sort
Filter
Clear
Reapply
Advanced
Text 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
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 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
25
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
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 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
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

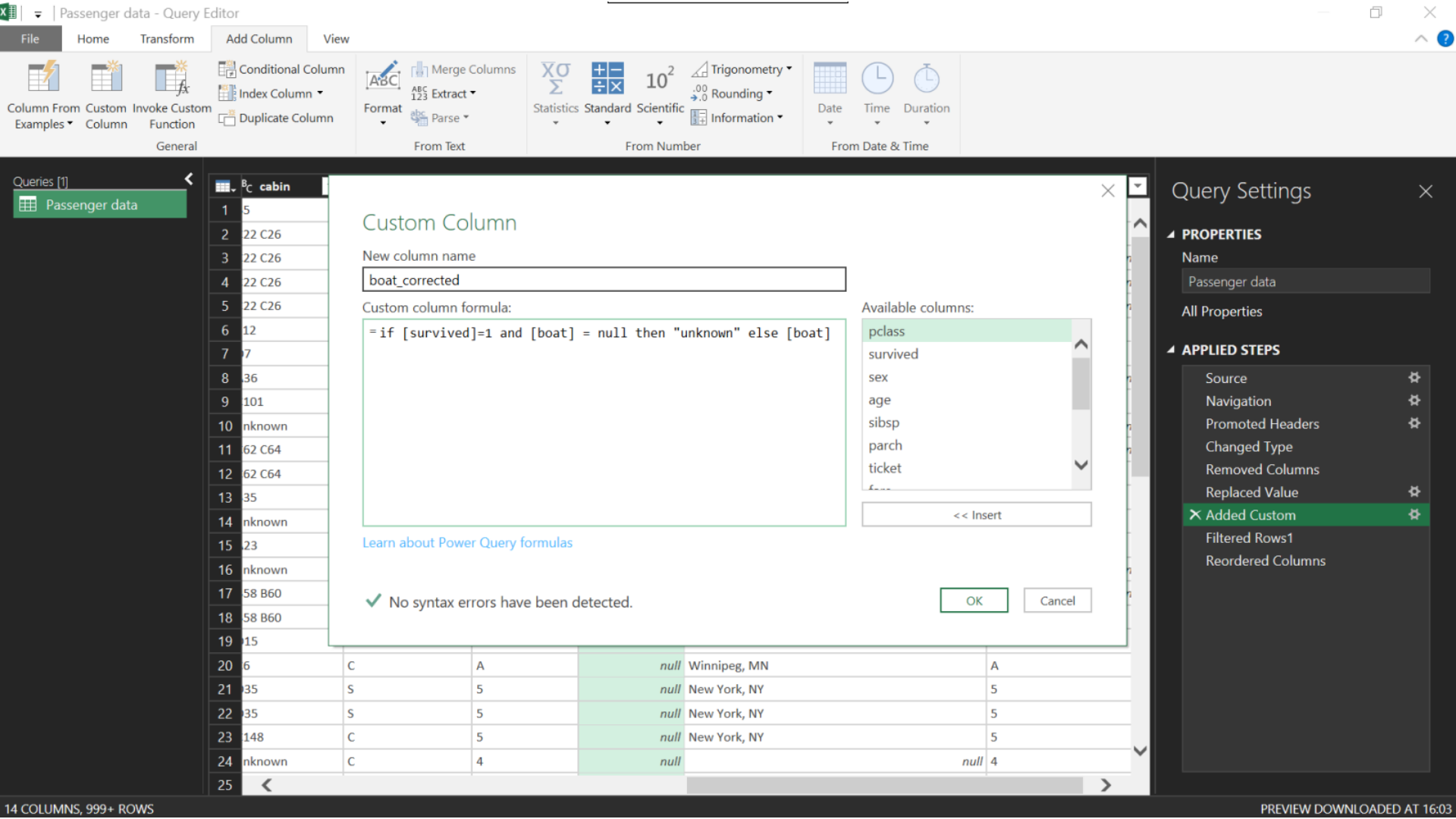
Cleansing data

- 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:

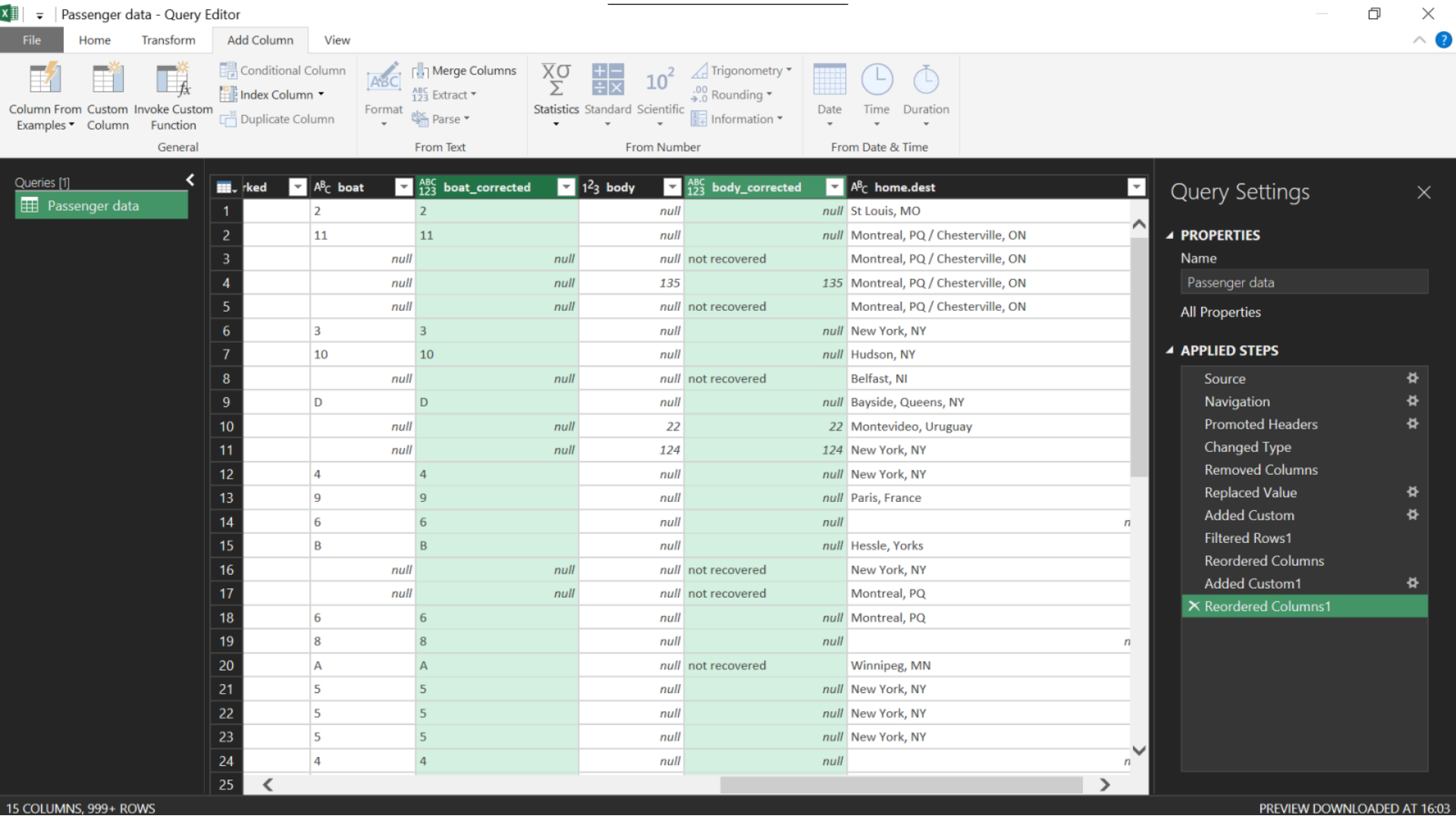
```
if [survived]=1 and [boat] = null then "unknown" else [boat]
```

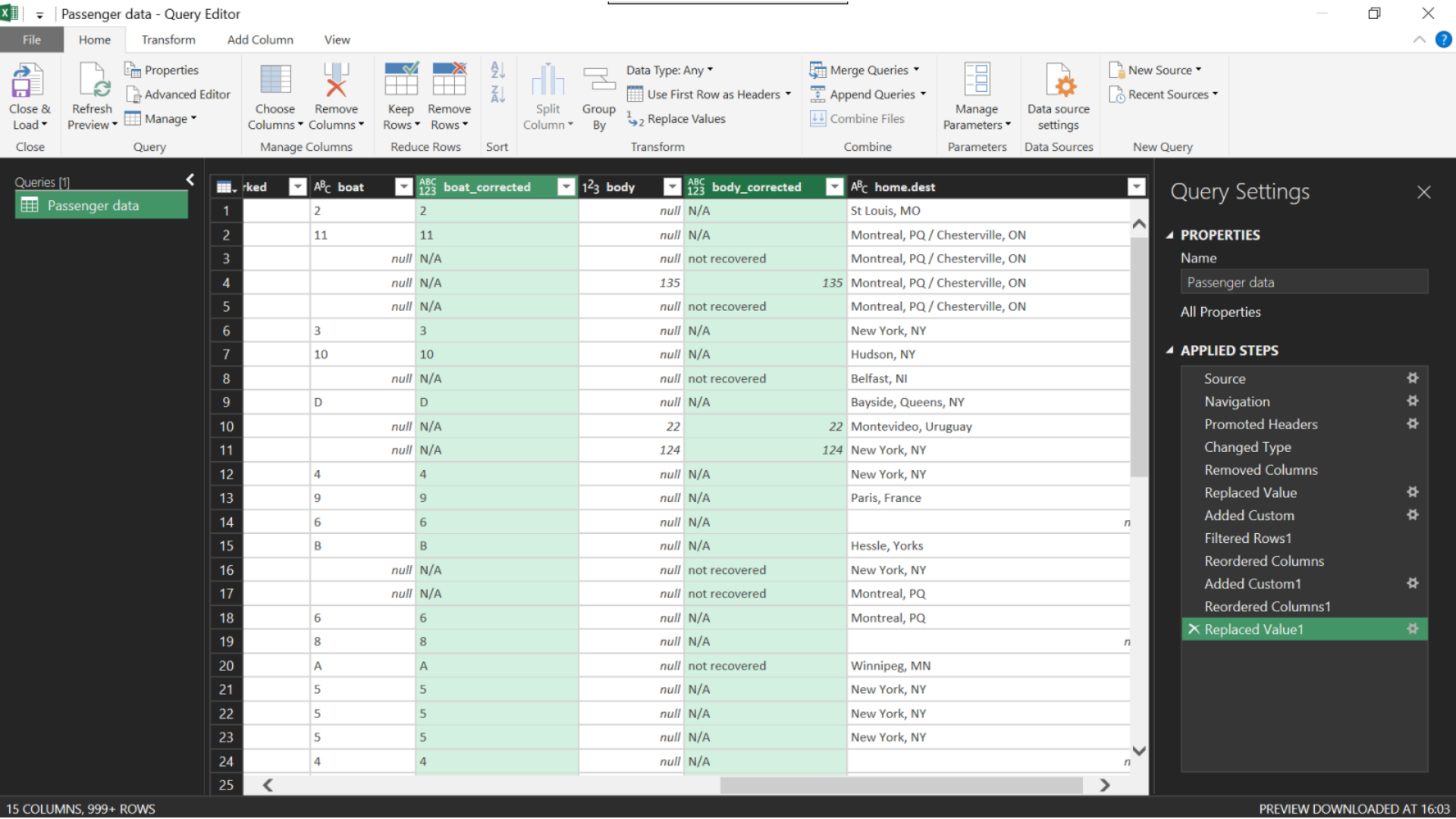

Passenger data - Query Editor
File
Home
Transform
Add Column
View
Column From Examples
Custom Column
Invoke Custom Function
Conditional Column
Index Column
Duplicate Column
General
Format
Merge Columns
Extract
Parse
From Text
Statistics
Standard
Scientific
Trigonometry
Rounding
Information
From Number
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
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Any
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 [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

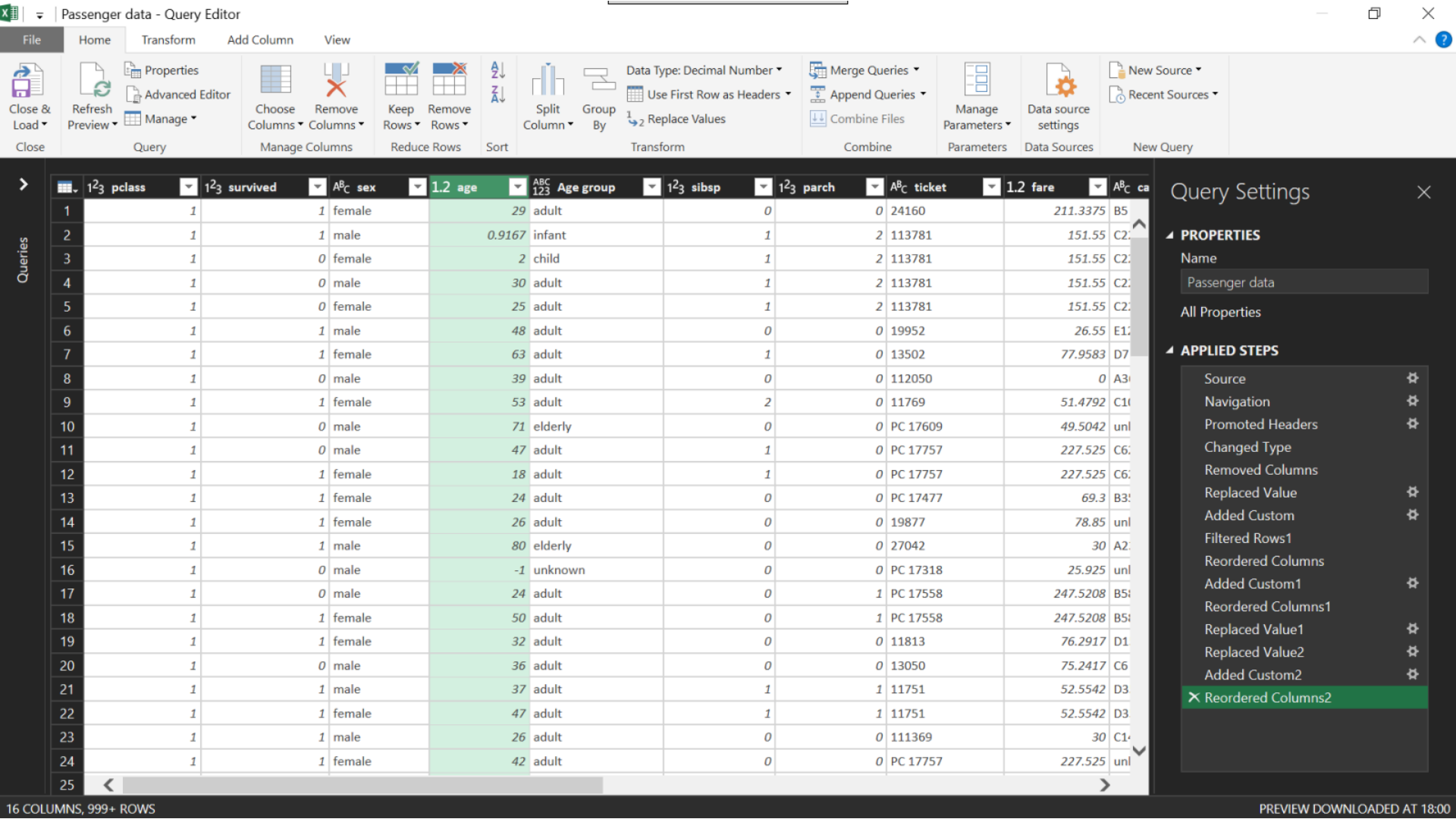
	Column Name	Operator	Value		Output	
If	age	equals	-1	Then	unknown	...
Else If	age	is less than	1	Then	infant	
Else If	age	is less than	12	Then	child	
Else If	age	is less than	18	Then	teenager	
Else If	age	is less than	65	Then	adult	
Else If	age	is greater than or...	65	Then	elderly	

Add rule

Otherwise

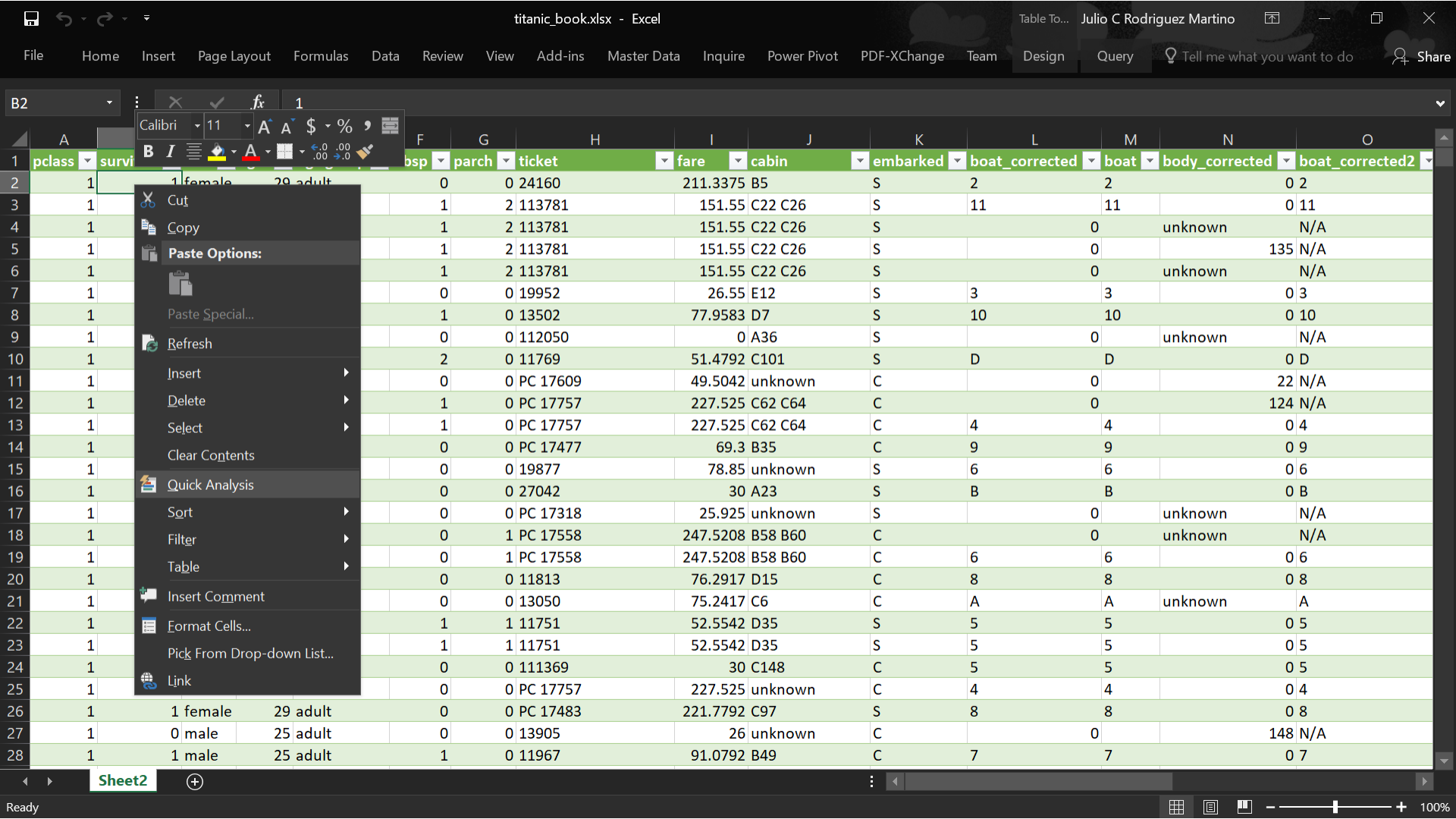
unknown

OK Cancel


Passenger data - Query Editor
File
Home
Transform
Add Column
View
Close & Load
Refresh Preview
Properties
Advanced Editor
Manage
Close
Query
Choose Columns
Remove Columns
Manage Columns
Keep Rows
Remove Rows
Reduce Rows
Sort
Split Column
Group By
Data Type: Decimal Number
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
Age group
sibsp
parch
ticket
fare
ca
1 1 female 29 adult 0 0 24160 211.3375 B5
2 1 1 male 0.9167 infant 1 2 113781 151.55 C2
3 1 0 female 2 child 1 2 113781 151.55 C2
4 1 0 male 30 adult 1 2 113781 151.55 C2
5 1 0 female 25 adult 1 2 113781 151.55 C2
6 1 1 male 48 adult 0 0 19952 26.55 E1
7 1 1 female 63 adult 1 0 13502 77.9583 D7
8 1 0 male 39 adult 0 0 112050 0 A3
9 1 1 female 53 adult 2 0 11769 51.4792 C1
10 1 0 male 71 elderly 0 0 PC 17609 49.5042 unl
11 1 0 male 47 adult 1 0 PC 17757 227.525 C6
12 1 1 female 18 adult 1 0 PC 17757 227.525 C6
13 1 1 female 24 adult 0 0 PC 17477 69.3 B3
14 1 1 female 26 adult 0 0 19877 78.85 unl
15 1 1 male 80 elderly 0 0 27042 30 A2
16 1 0 male -1 unknown 0 0 PC 17318 25.925 unl
17 1 0 male 24 adult 0 1 PC 17558 247.5208 B5
18 1 1 female 50 adult 0 1 PC 17558 247.5208 B5
19 1 1 female 32 adult 0 0 11813 76.2917 D1
20 1 0 male 36 adult 0 0 13050 75.2417 C6
21 1 1 male 37 adult 1 1 11751 52.5542 D3
22 1 1 female 47 adult 1 1 11751 52.5542 D3
23 1 1 male 26 adult 0 0 111369 30 C1
24 1 1 female 42 adult 0 0 PC 17757 227.525 unl
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
Replaced Value2
Added Custom2
Reordered Columns2
16 COLUMNS, 999+ ROWS
PREVIEW DOWNLOADED AT 18:00

Visualizing data for preliminary analysis

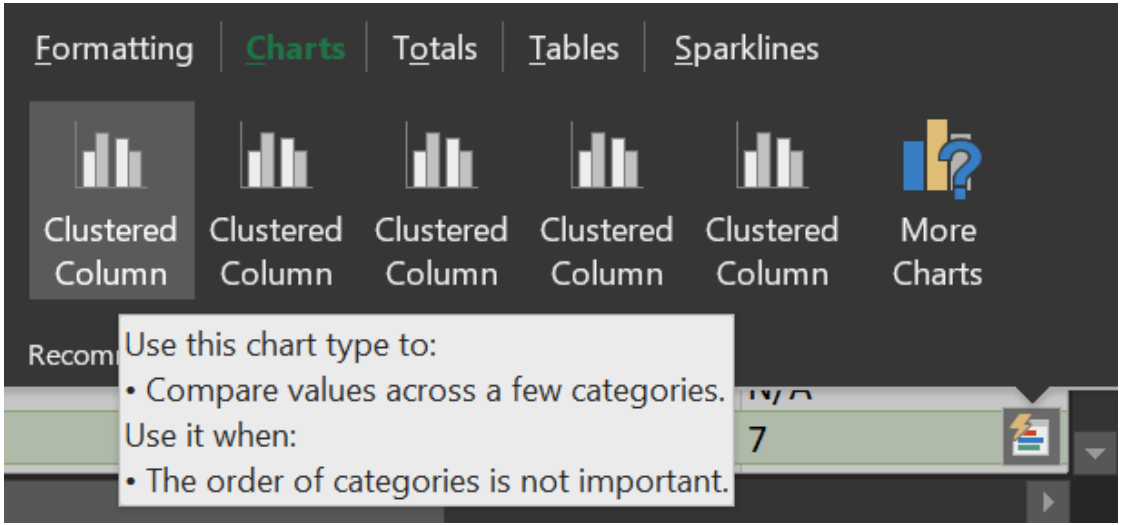
- After cleaning the dataset, it is always recommended to visualize it.
- This helps us gain an understanding of the different variables, how their values are distributed, and the correlations that exist between them (we will explore correlations in more detail in the next lesson).
- We can determine 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:



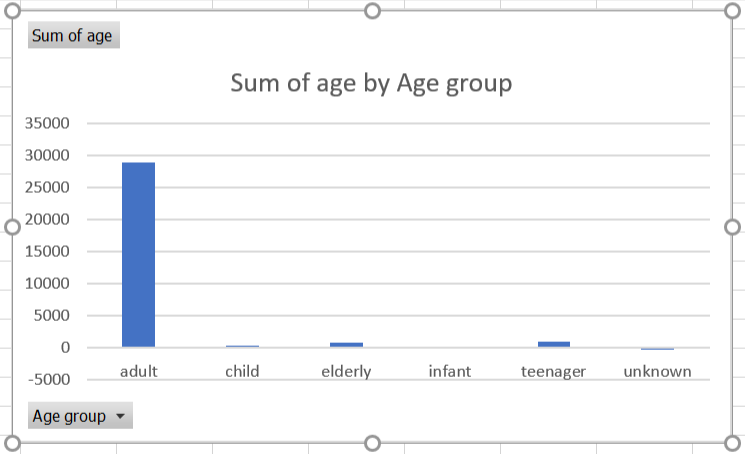
titanic_book.xlsx - Excel

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

Drag fields between areas b

Filters

Axis (Categories)

Age group

Sum of age

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...

Defer Layout Update

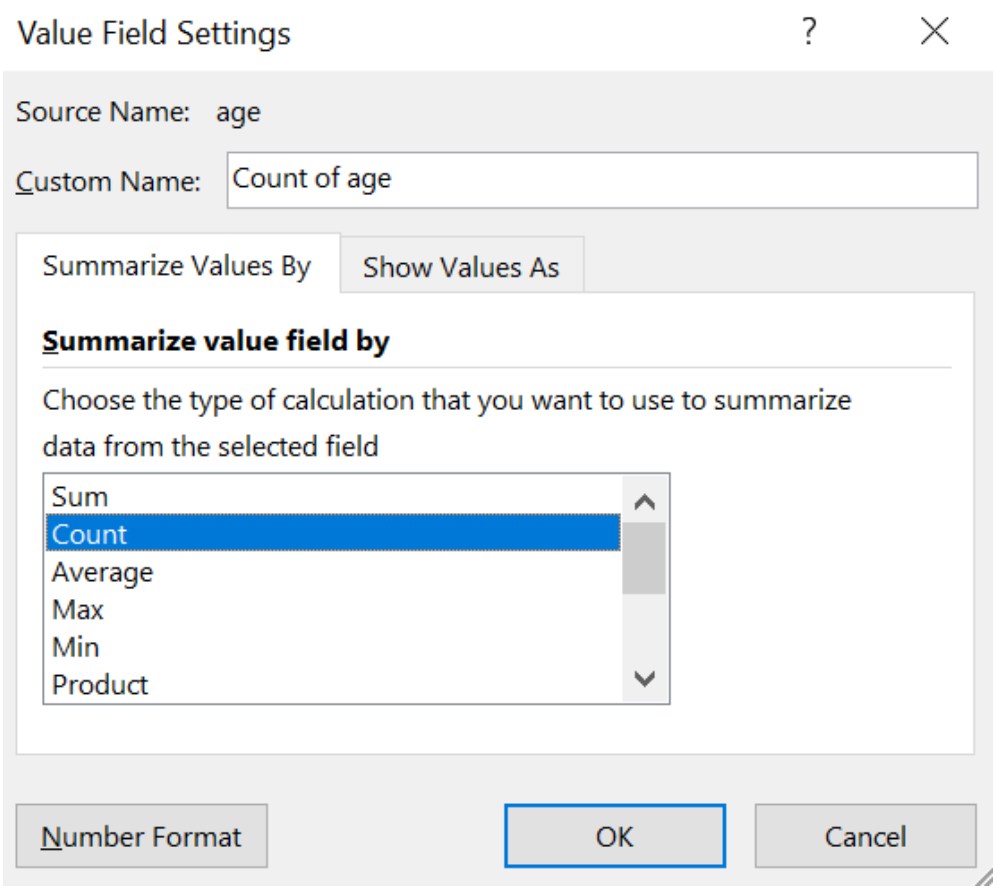
Update

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:



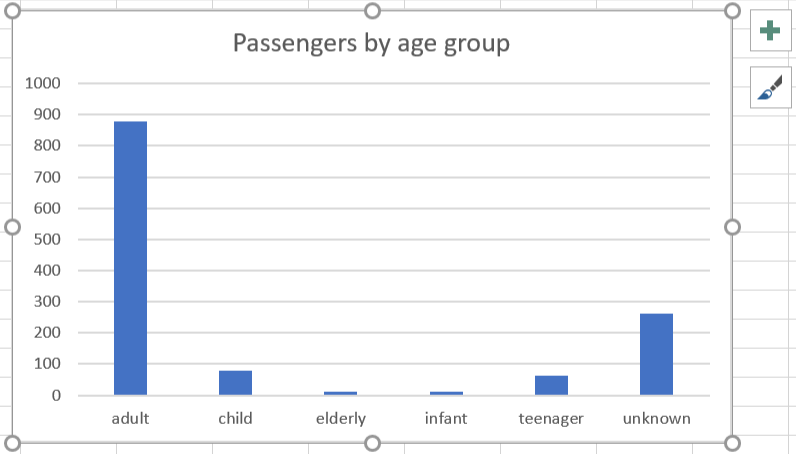
titanic_book.xlsx - Excel

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

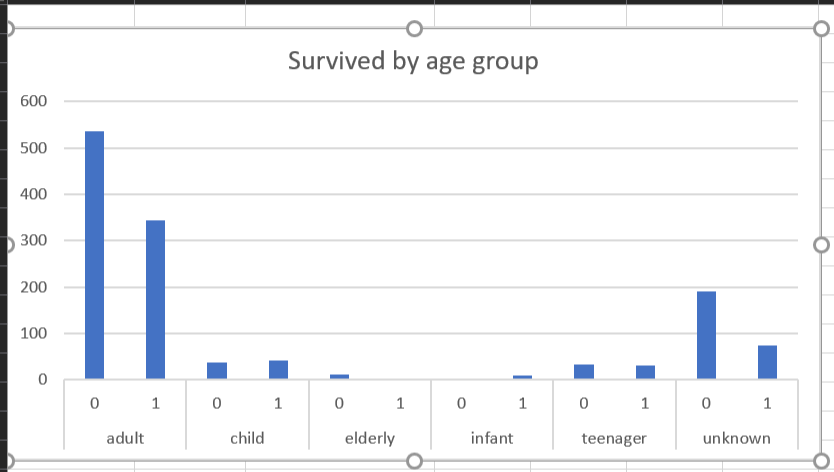
100%

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)	Values
Age group	Count of Age group
survived	

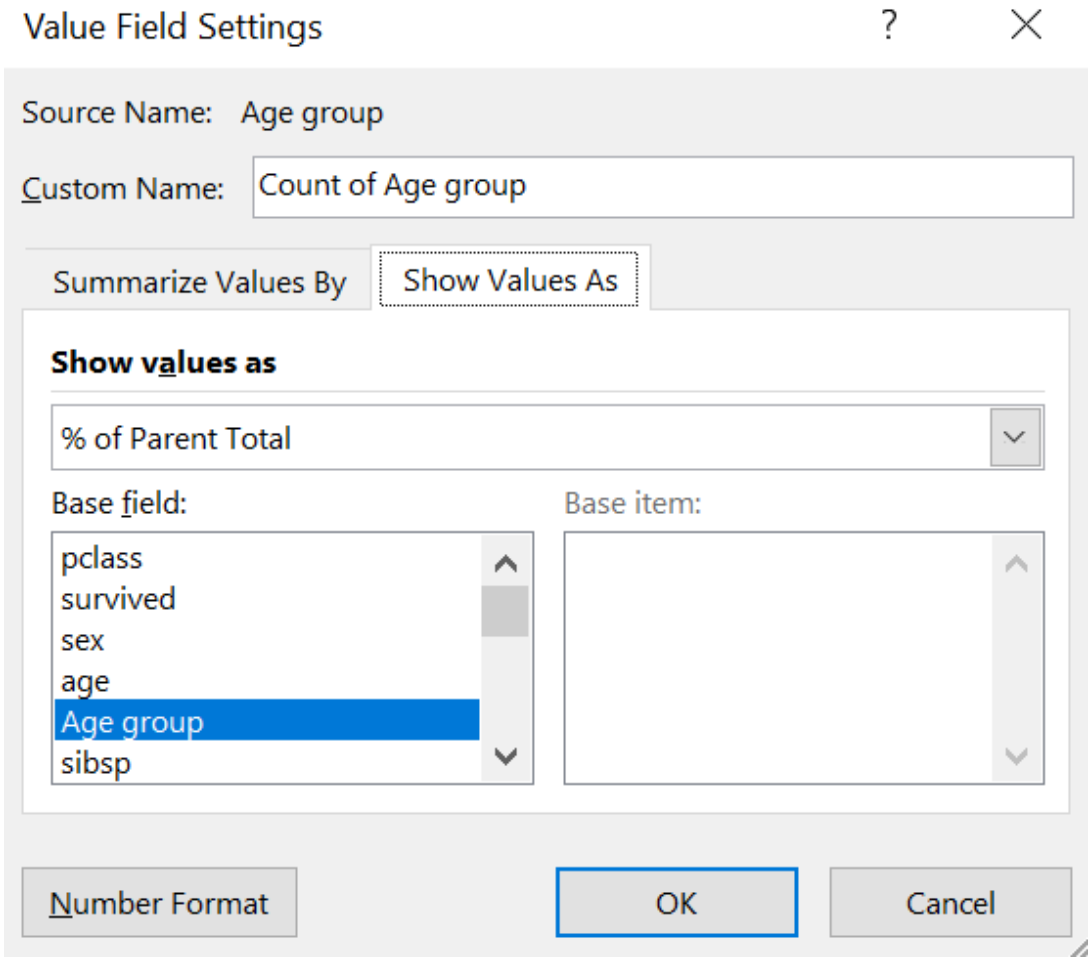
Defer Layout Update

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

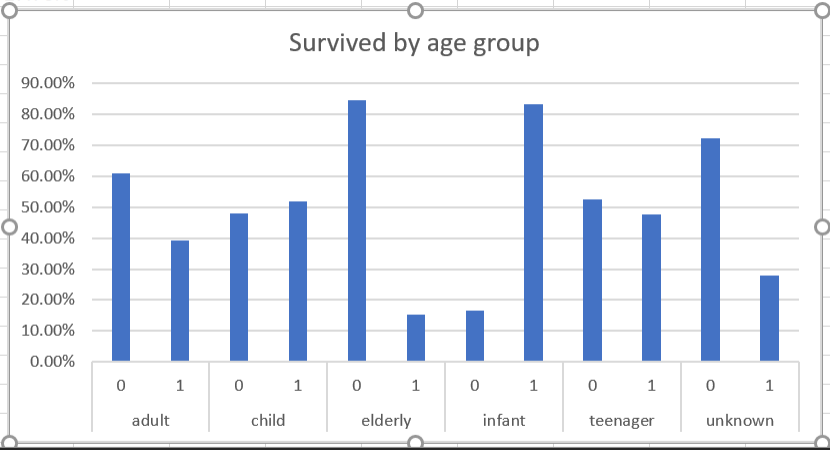
titanic_book.xlsx - Excel

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%



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%

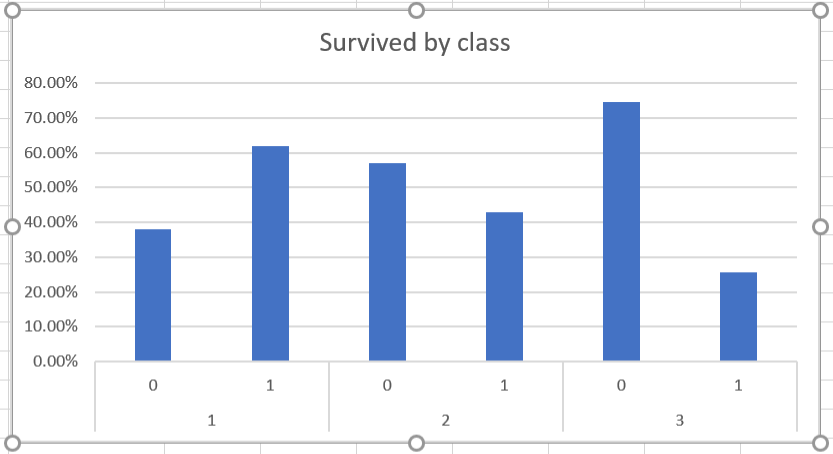
titanic_book.xlsx - Excel

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

PivotChart Fields

Choose fields to add to report:

Search

- [x] 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

Ready 100%

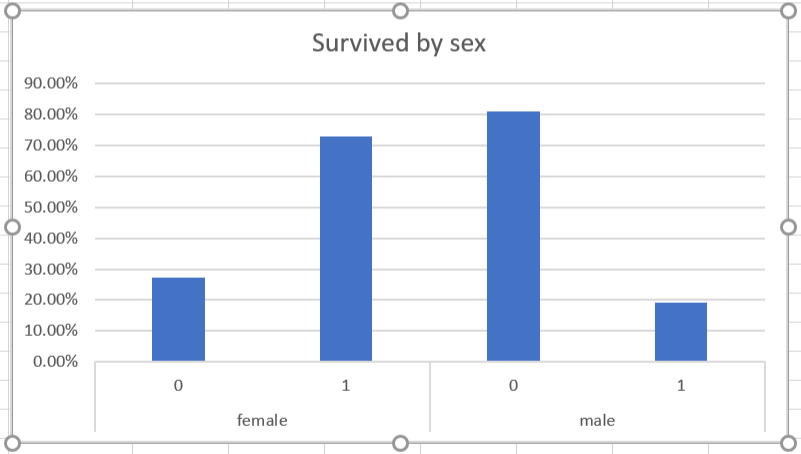
titanic_book.xlsx - Excel

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

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

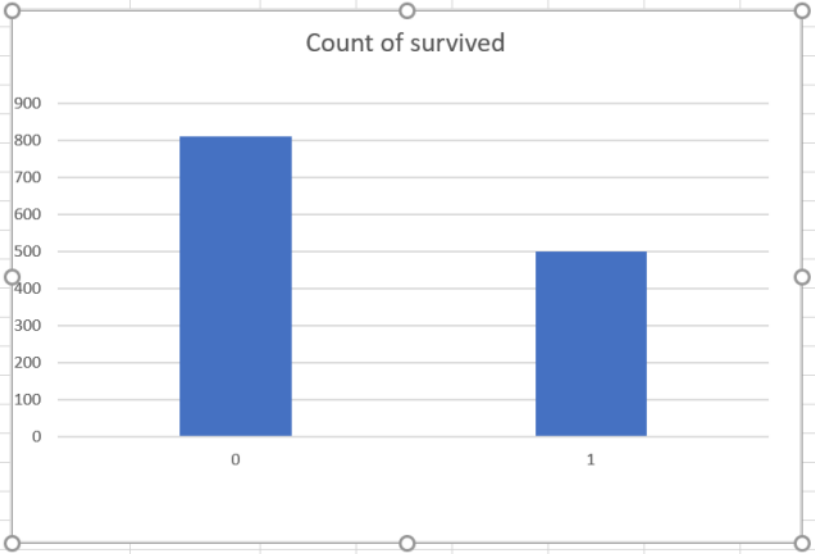
titanic_book.xlsx - Excel

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

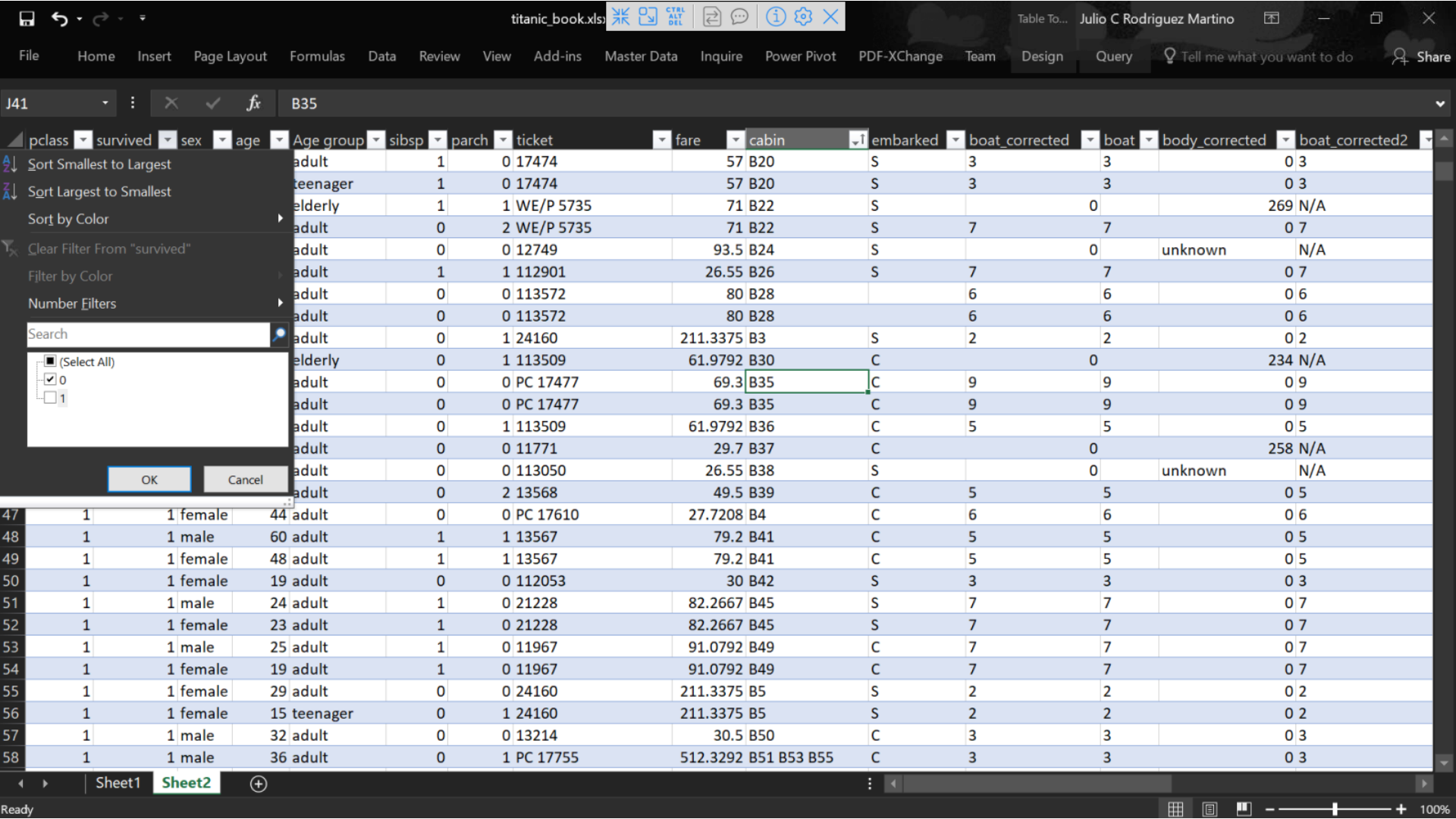
Sheet1 Sheet2

Ready

100%

Understanding unbalanced datasets

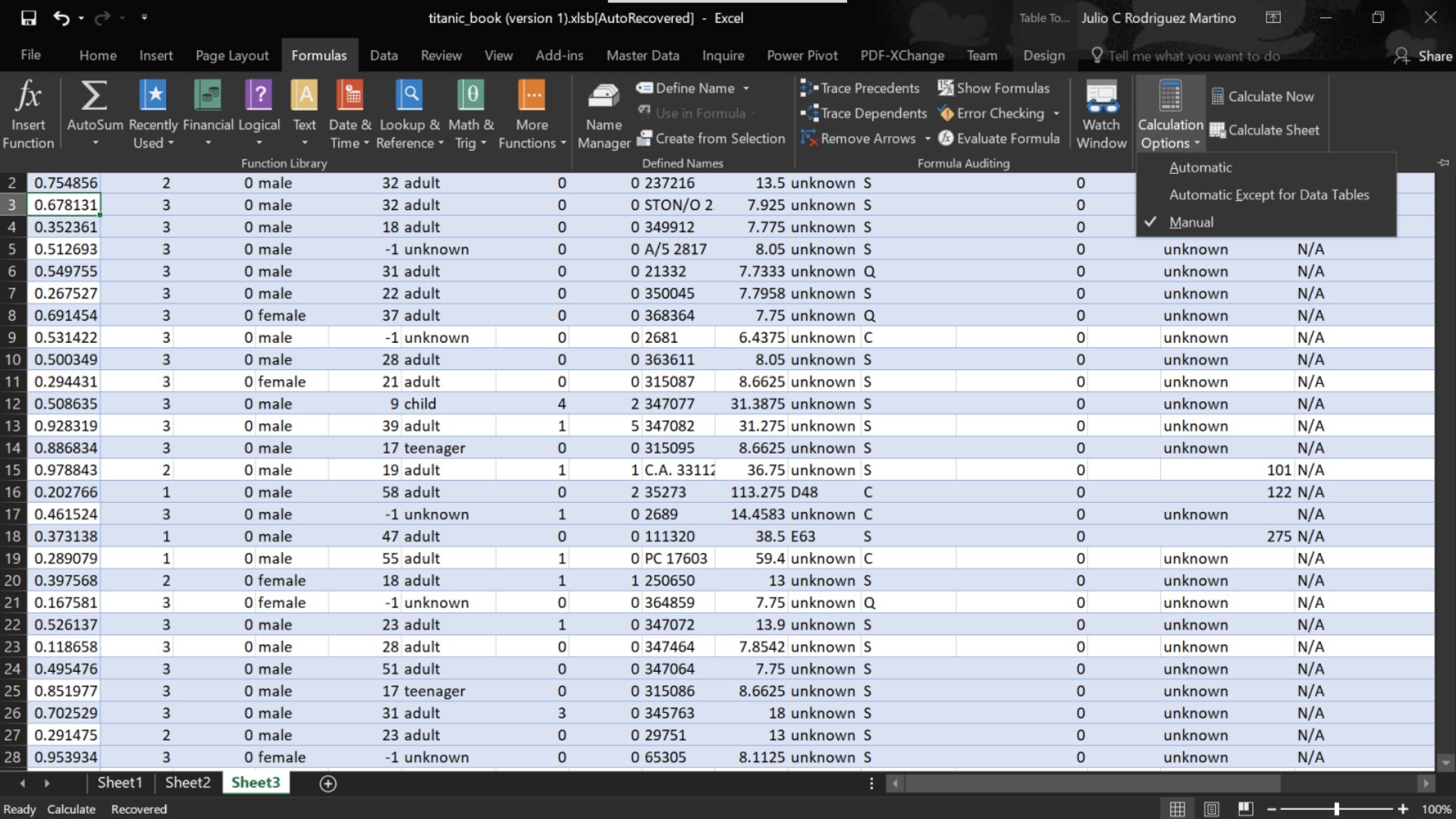
- It is clear from the preceding diagram and table that there are nearly twice as many non-survivors than survivors.
- If we use this dataset as is, we are introducing a bias to our dataset that will affect the results.
- Predicting 0 for the survival variable will be approximately two times more probable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


titanic_book.xlsx
Julio C Rodriguez Martino
File Home Insert Page Layout Formulas Data Review View Add-ins Master Data Inquire Power Pivot PDF-XChange Team Design Query Tell me what you want to do Share
J41
B35
pclass survived sex age Age group sibsp parch ticket fare cabin embarked boat_corrected boat body_corrected boat_corrected2
Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Clear Filter From "survived"
Filter by Color
Number Filters
Search
(Select All)
0
1
OK
Cancel
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
Sheet1 Sheet2
Ready
100%

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()



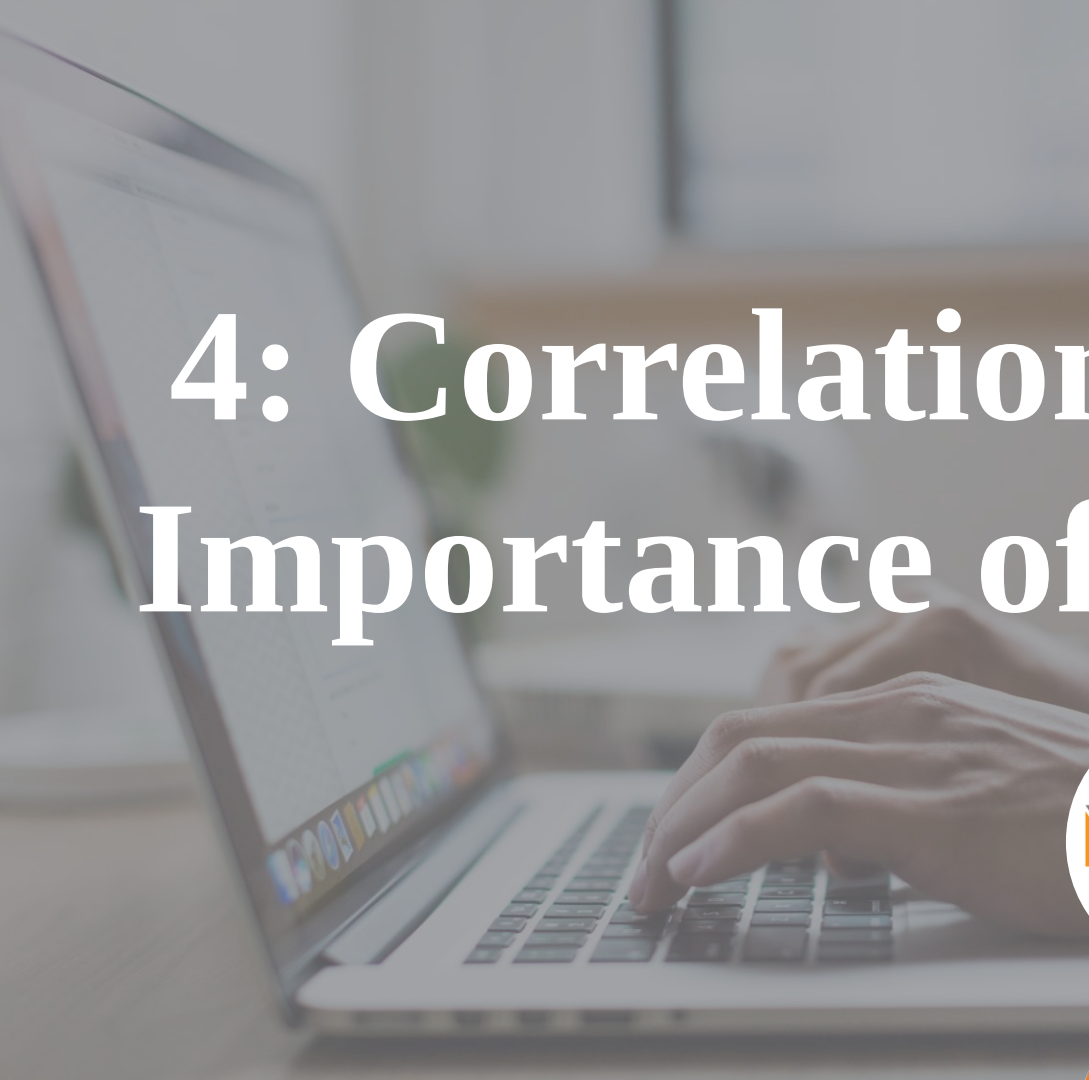
2	0.754856	2	0	male	32	adult	0	0	237216	13.5	unknown	S	0	
3	0.678131	3	0	male	32	adult	0	0	STON/O 2	7.925	unknown	S	0	
4	0.352361	3	0	male	18	adult	0	0	349912	7.775	unknown	S	0	
5	0.512693	3	0	male	-1	unknown	0	0	A/5 2817	8.05	unknown	S	0	unknown N/A
6	0.549755	3	0	male	31	adult	0	0	21332	7.7333	unknown	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. 33112	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



4: 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

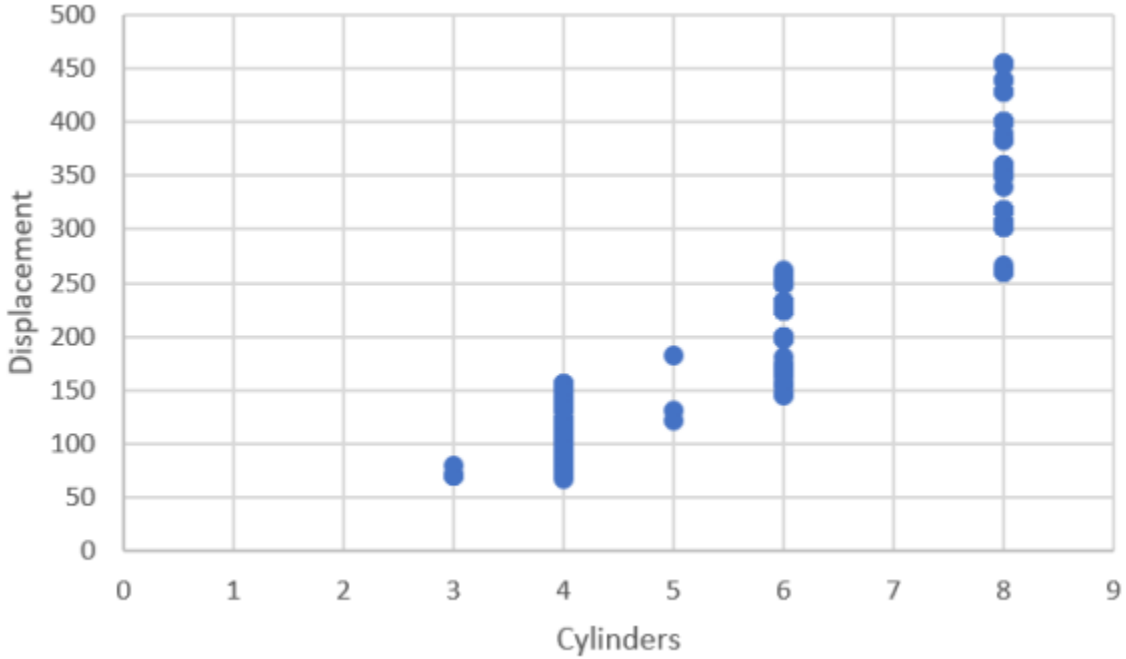
- You will need to download the auto-mpg.xlsx file from the GitHub.

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.

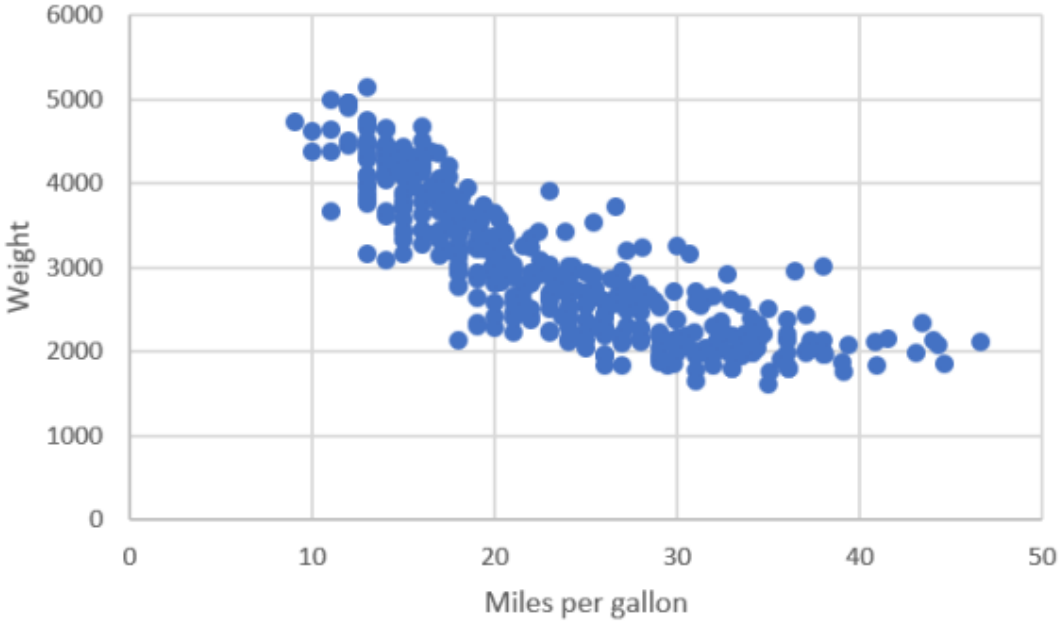
Building a scatter diagram

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

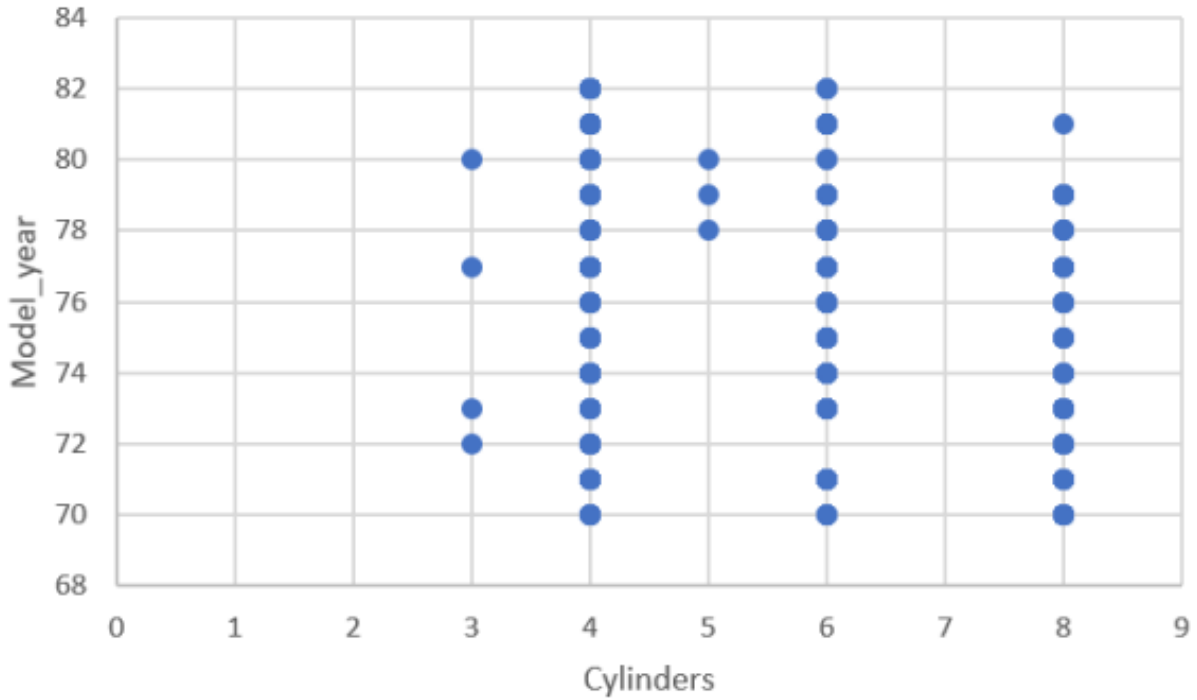


Building a scatter diagram

- 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:

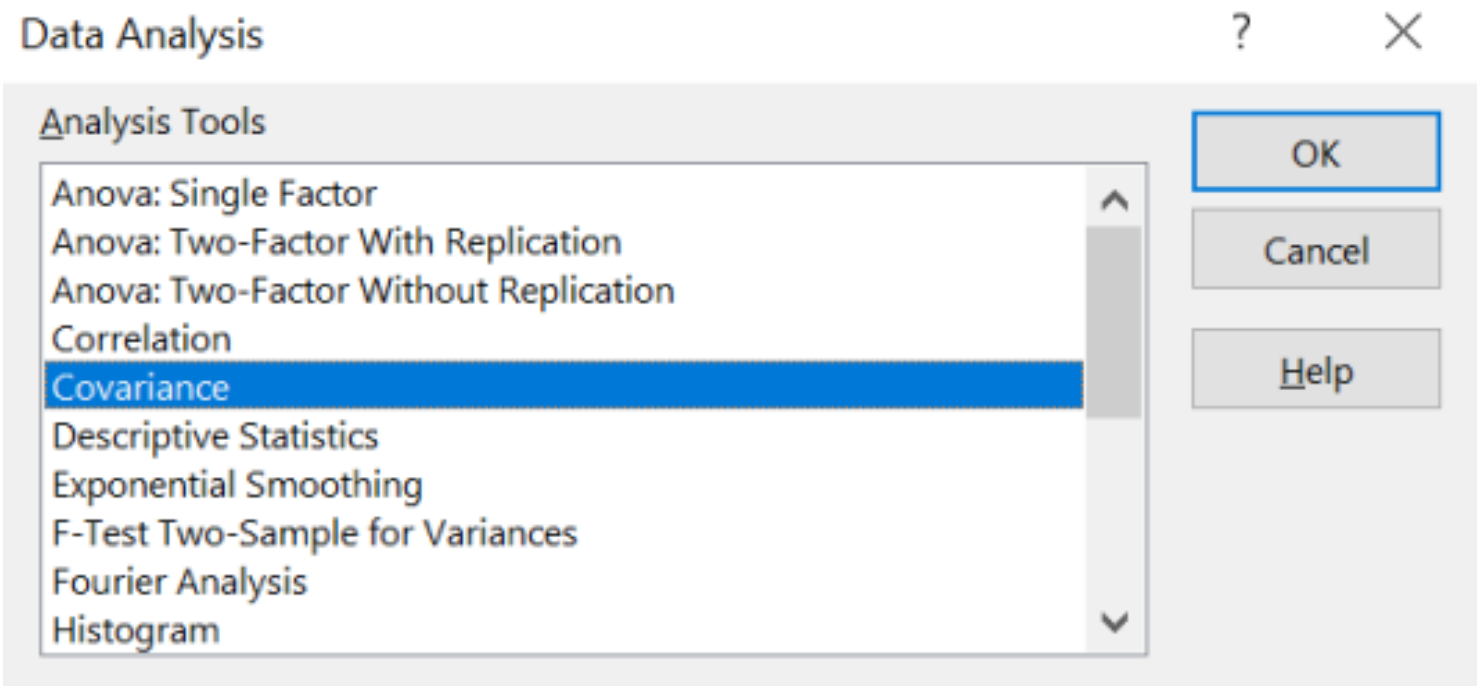
$$Cov(x, y) = mean[(x - \hat{x}).(y - \hat{y})]$$

Calculating the covariance

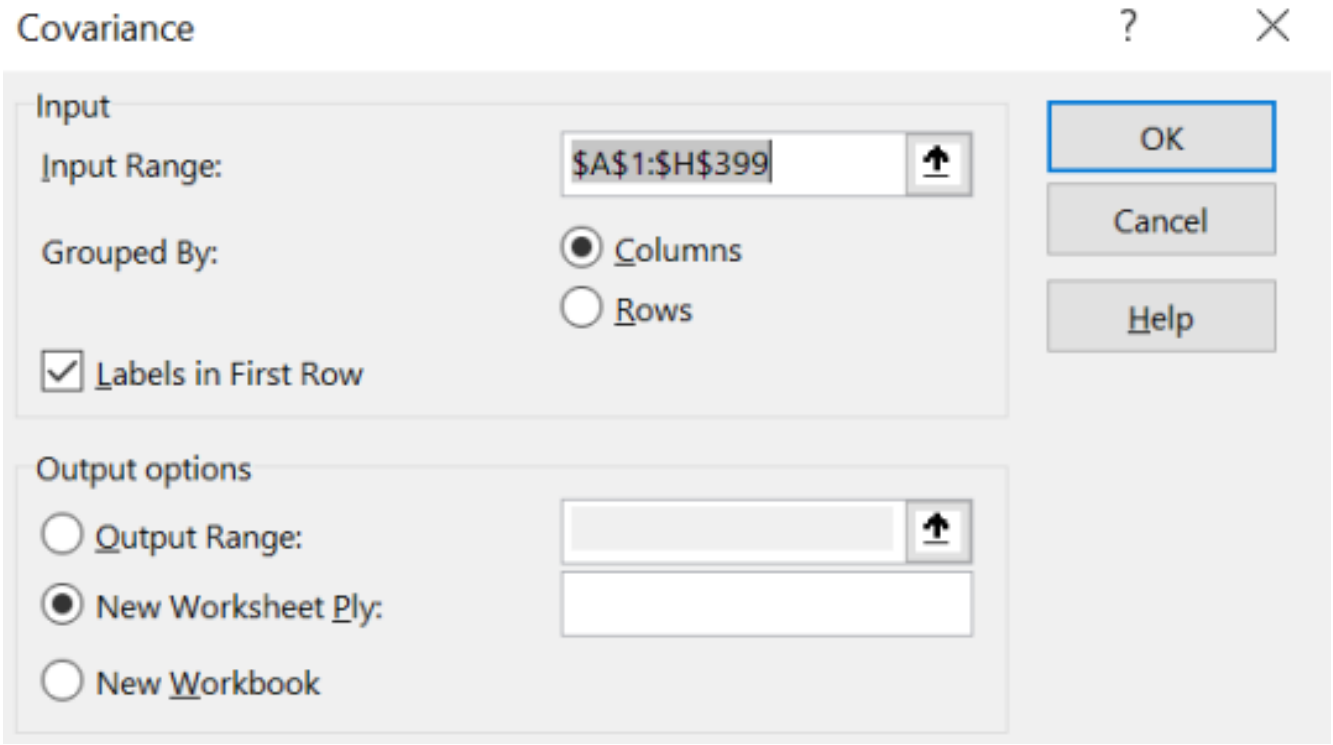
To calculate the covariances, perform the following steps:

- 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])

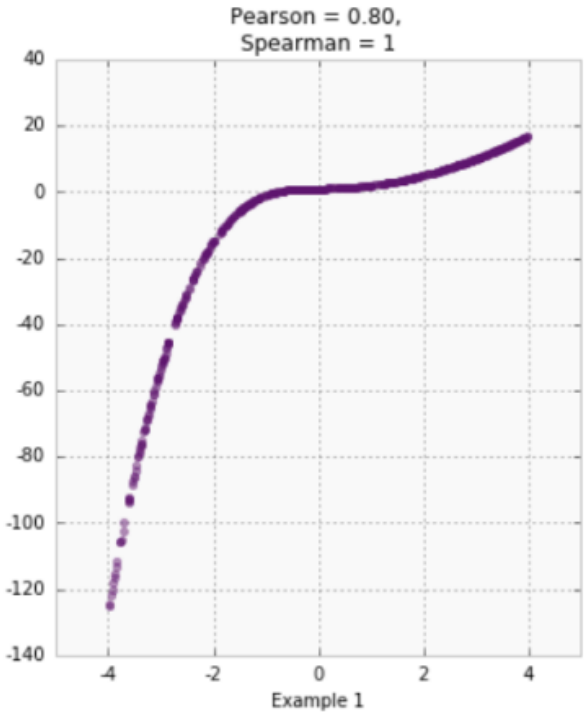
	A	B	C	D	E	F	G	H
1	Rank_mpg	Rank_cylinders	Rank_displacement	Rank_horsepower	Rank_weight	Rank_acceleration	Rank_model_year	Rank_origin
2	283	52	75	94	109	362.5	384	274
3	337.5	52	46.5	35.5	90	372	384	274
4	283	52	65	56.5	115	384	384	274
5	318	52	84	56.5	116	362.5	384	274
6	303	52	93	81	112	388	384	274
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

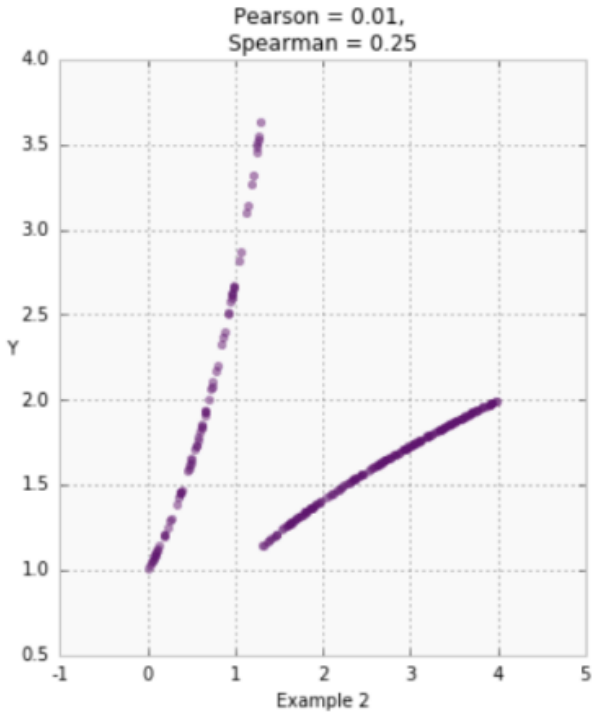
... Data | Covariance | Pearson | **Ranks** | Spearman | Sheet8

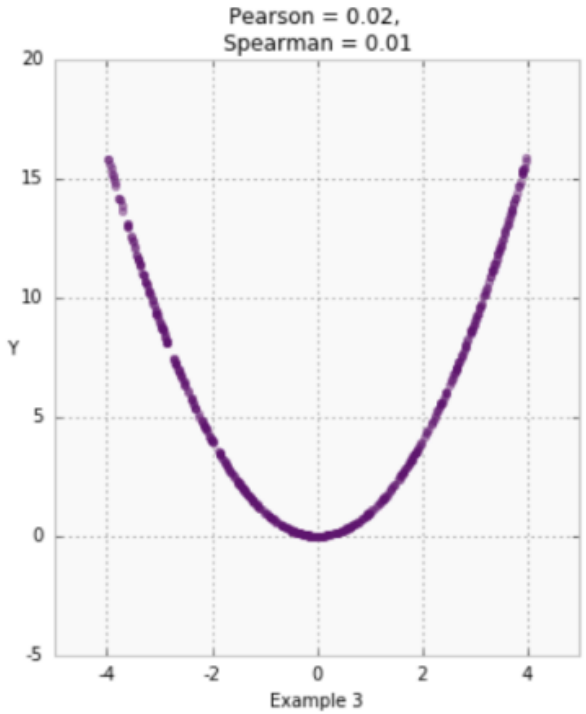
- Because horsepower is missing some values, they cannot be ranked and so appear as #N/A. Since there are only a few of them, we can remove them manually.
- This will avoid errors when calculating the Pearson coefficient in the next step, exactly as we did before; the result is as follows:

	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







Understanding least squares

- In this case, it is useful to rely on the least squares method. Given a set of points (xi,yi) and a function such as y'i = f(xi), this method minimizes the square of the differences between y'i and yi.
- The general expression for the minimization that we are calculating is as follows:

$$min(\sum_{i}(y' - y)^2)$$

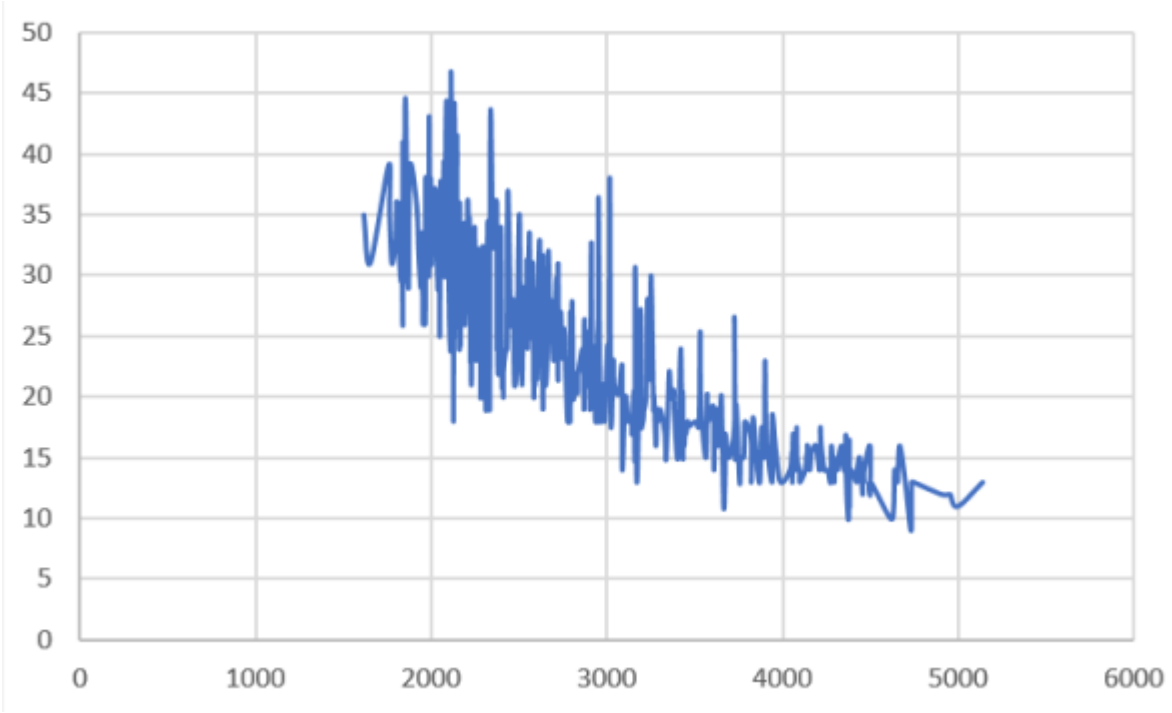
Understanding least squares

We will use two columns from our data table, namely weight and mpg:

- Create a new table in a new sheet.
- Copy the values of the weight and mpg columns.
- Order the rows by the value of weight; the resulting table is as follows:

	A	B
1	weight	mpg
2	1613	35
3	1649	31
4	1755	39.1
5	1760	35.1
6	1773	31
7	1795	33
8	1795	33
9	1800	36.1
10	1800	36.1
11	1825	29.5
12	1825	36
13	1834	27
14	1835	26
15	1835	40.9
16	1836	32
17	1845	29.8
18	1850	44.6
19	1867	29
20	1875	39
21	1915	35.7
22	1925	31.9
23	1937	29

Understanding least squares



Understanding least squares

- Create a new column, prediction, using the following formula:

=$H$2*POWER([@weight];$H$3)

- The resulting table is as follows:

	A	B	C
1	weight	mpg	predict
2	1613	35	1.493943
3	1649	31	1.477546
4	1755	39.1	1.43223
5	1760	35.1	1.430194
6	1773	31	1.424941
7	1795	33	1.416182
8	1795	33	1.416182
9	1800	36.1	1.414214
10	1800	36.1	1.414214
11	1825	29.5	1.404494
12	1825	36	1.404494
13	1834	27	1.401043
14	1835	26	1.400662
15	1835	40.9	1.400662
16	1836	32	1.40028
17	1845	29.8	1.396861
18	1850	44.6	1.394972
19	1867	29	1.388606
20	1875	39	1.385641

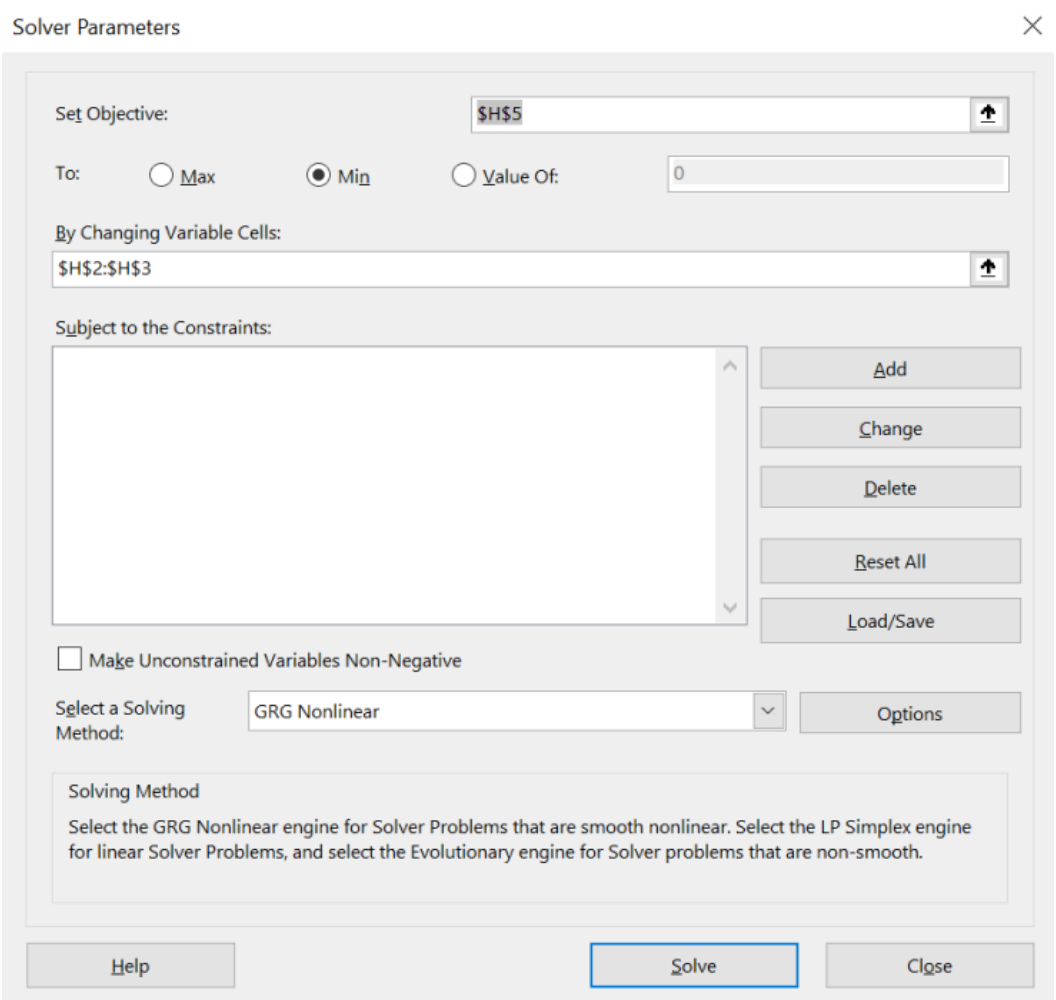
Understanding least squares

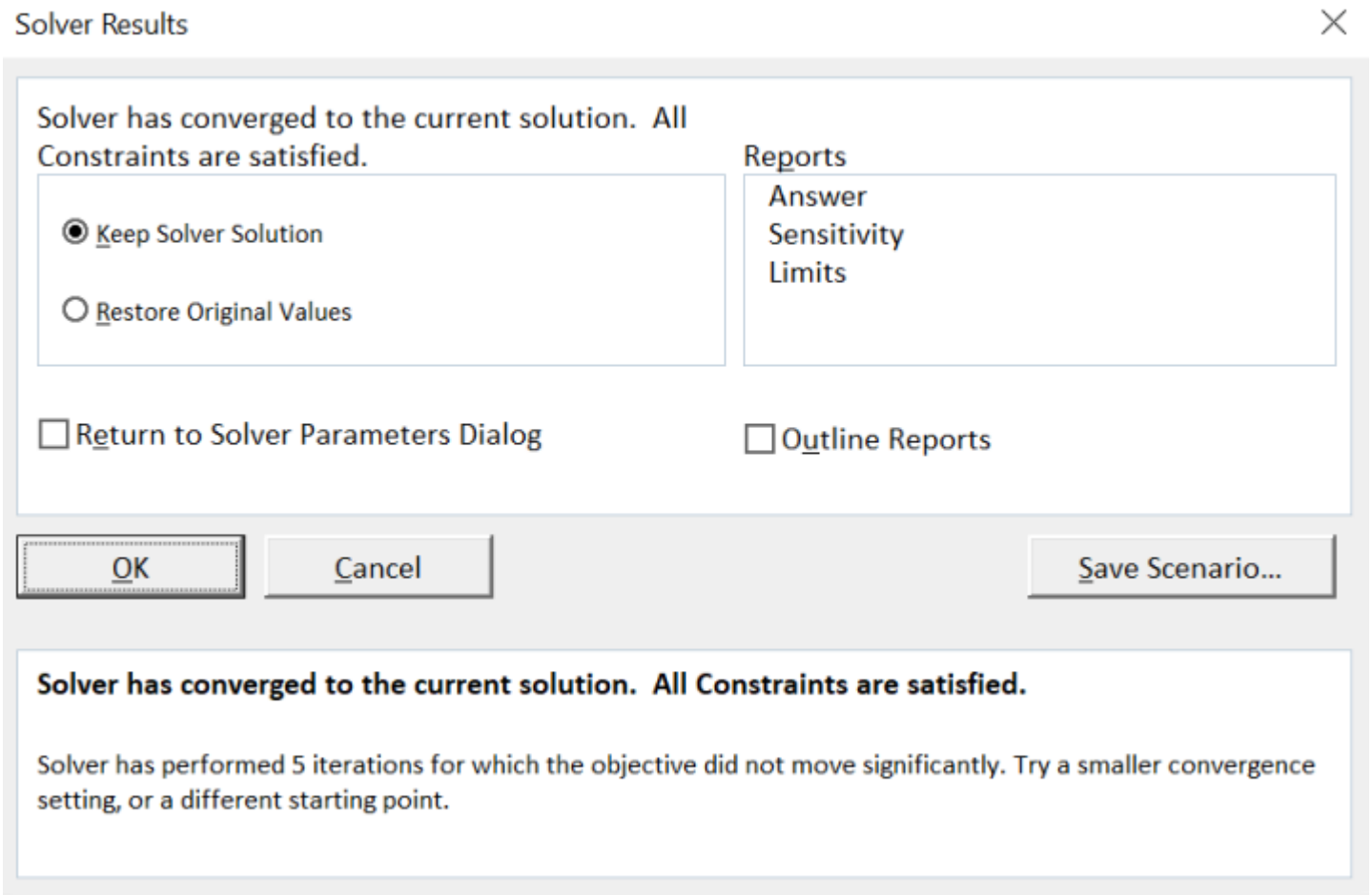
- The quantity to minimize is the sum of the squares of the errors. To calculate it, we create a new column, Squared error, with the following formula:

=([@mpg]-[@prediction])^2

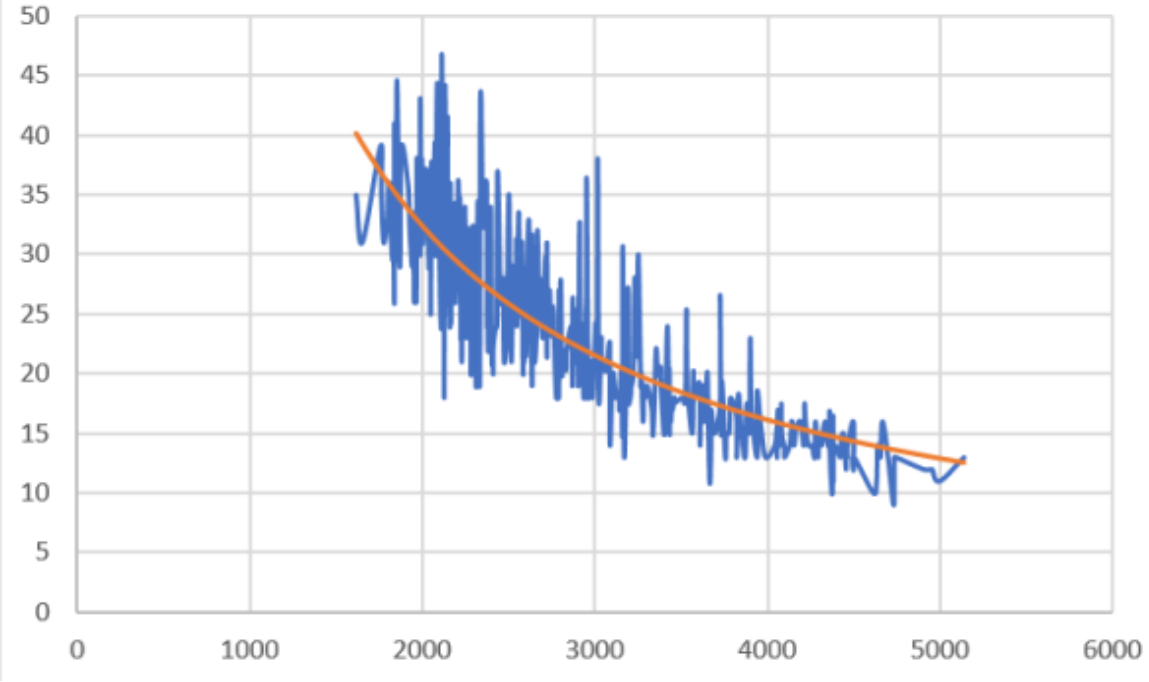
- Then, use the following formula to sum all the values in that column in a cell:

=SUM(Table9[Squared error])


Solver Parameters
Set Objective:
$H$5
To:
Max
Min
Value Of:
0
By Changing Variable Cells:
$H$2:$H$3
Subject to the Constraints:
Add
Change
Delete
Reset All
Load/Save
Make Unconstrained Variables Non-Negative
Select a Solving Method:
GRG Nonlinear
Options
Solving Method
Select the GRG Nonlinear engine for Solver Problems that are smooth nonlinear. Select the LP Simplex engine for linear Solver Problems, and select the Evolutionary engine for Solver problems that are non-smooth.
Help
Solve
Close


Solver Results
Solver has converged to the current solution. All Constraints are satisfied.
Keep Solver Solution
Restore Original Values
Reports
Answer
Sensitivity
Limits
Return to Solver Parameters Dialog
Outline Reports
OK
Cancel
Save Scenario...
Solver has converged to the current solution. All Constraints are satisfied.
Solver has performed 5 iterations for which the objective did not move significantly. Try a smaller convergence setting, or a different starting point.

	a	68563.9126					
	b	-1.0074493					
	SSE	7117.53396					



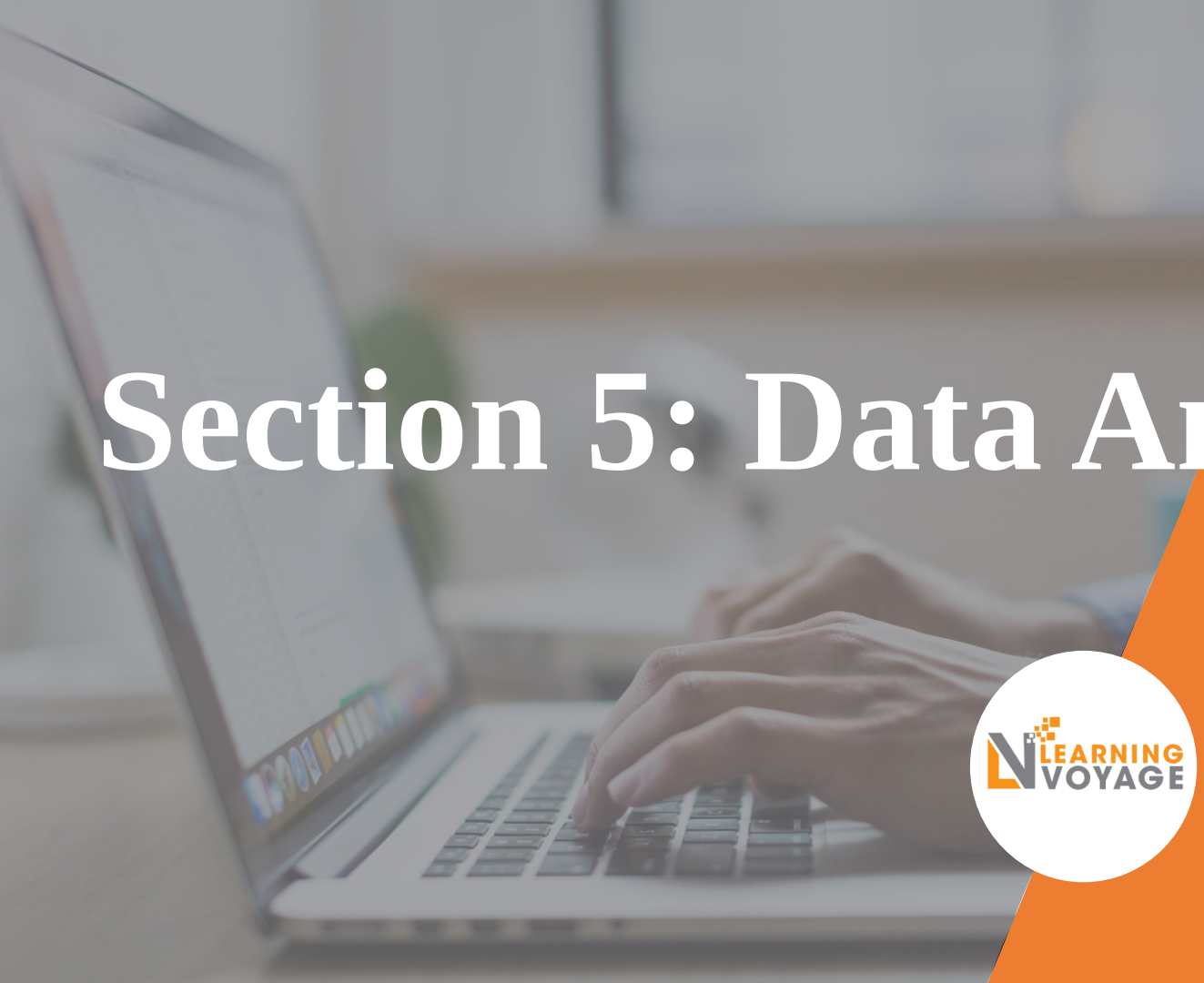
Focusing on feature selection

- There are automatic techniques to perform feature engineering, which are part of what is generically called Automatic Machine Learning (AutoML).
- The method consists of letting the computer try different feature sets, including combinations of them, and test the results until the best set is found.
- In spite of this, there is no general recipe for selecting features, and each problem has to be analyzed—in particular, finding the set of features that lead to a better model training and predictive power.

Summary

- In this lesson, we described the most widely used methods to establish correlations between variables, which will later be used as features in a machine learning model, This is a long and difficult task, but is the basis of a good predictive model.
- No method can be used alone to determine which features are important and which can be discarded.
- A combination of methods, plus a deep knowledge of the dataset, are fundamental to complete this task.

Section 5: Data Analytics



Analytics and Machine Learning Models

This section comprises the following lessons:

- lesson 6, Data Mining Models in Excel Hands-On Examples
- lesson 7, Implementing Time Series

6: Data Mining Models in Excel Hands-On Examples

LEARNING VOYAGE

Data Mining Models in Excel Hands-On Examples

- Data mining is about finding hidden patterns and associations in data.
- A large number of analyses that can only be performed by a human in a reasonable time if the amount of data is small, can be done by a computer in a very short time.
- Before Excel 2016, it was possible to install an add-in (called Data Mining) that was packed with different methods and models that could be used, mostly as black boxes, to get insights and discover information in any dataset.

Data Mining Models in Excel Hands-On Examples

In this lesson, we will cover the following topics:

- Learning by example: Market Basket Analysis
- Learning by example: Customer Cohort Analysis

Technical requirements

- To complete this section, the reader will need to download the transactions_by_dept.csv and cohort_input_data.csv files from the GitHub repository

Learning by example – Market Basket Analysis

- We have all read the sentence in almost every online store: People who bought this product also bought.... It all started with Amazon in the 1990s, and it is widespread today.

- This same principle is even being tested in physical stores, where customers can get personalized suggestions on which items to buy based on their shopping history and similarity with other products.

- The resulting table looks like this:

	A	B	C	D
1	POS Txn	Dept	ID	Sales U
2	16120100160021008773	0261:HOSIERY	250	2
3	16120100160021008773	0634:VITAMINS & HLTH AIDS	102	1
4	16120100160021008773	0879:PET SUPPLIES	158	2
5	16120100160021008773	0973:CANDY	175	2
6	16120100160021008773	0982:SPIRITS	176	1
7	16120100160021008773	0983:WINE	177	4
8	16120100160021008773	0991:TOBACCO	179	2
9	16120100160021008774	0597:HEALTH AIDS	93	1
10	16120100160021008774	0604:PERSONAL CARE	100	5
11	16120100160021008775	0819:PRE-RECORDED A/V	135	1
12	16120100160021008775	0826:SMALL ELECTRICS	138	1
13	16120100160021008775	0982:SPIRITS	176	1
14	16120100160021008776	0961:GENERAL GROCERIES	169	3
15	16120100160021008777	0982:SPIRITS	176	2
16	16120100160021008778	0982:SPIRITS	176	4
17	16120100160021008778	0991:TOBACCO	179	1
18	16120100160021008779	0879:PET SUPPLIES	158	16
19	16120100160021008779	0982:SPIRITS	176	1
20	16120100160021008779	0983:WINE	177	2
21	16120100160021008779	0984:BEER	178	1
22	16120100160021008780	0530:SCHOOL/OFFIC SUPP	70	1
23	16120100160021008780	0597:HEALTH AIDS	93	1
24	16120100160021008780	0601:VALUE ZONE	97	1
25	16120100160021008780	0634:VITAMINS & HLTH AIDS	102	1

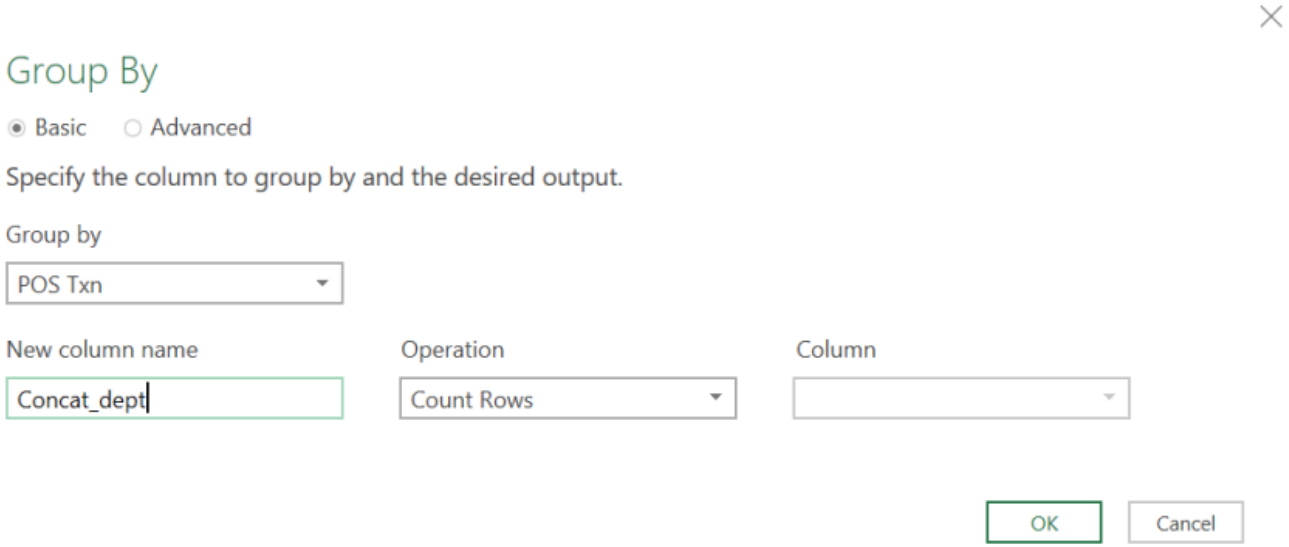
Learning by example – Market Basket Analysis

To start our analysis, we need to do the following:

- Group all transactions by transaction number.
- Build a list of all the departments visited in each particular purchase.

To perform these tasks, we will use the Power Query capabilities, following these steps:

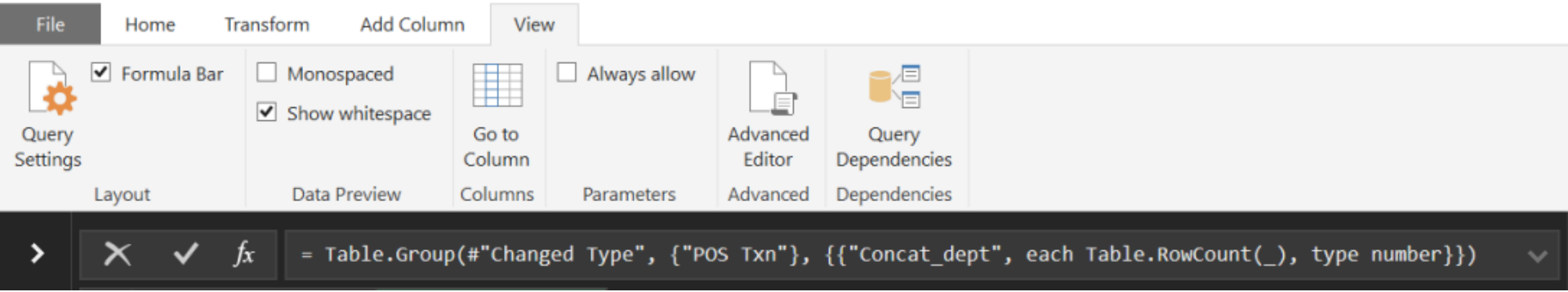
- Navigate to Data | From Table/Range.
- Make sure that the data type of the POS Txn column is set to Text before continuing.
- Click on Group by. You will see the following pop-up window:



- Change the calculation formula manually to get the data transformation we want. Navigate to View and make sure that Formula Bar is selected. If not, select it, and you will see the following formula:

```
= Table.Group(#"Changed Type", {"POS Txn"}, {{"Concat_dept", each Table.RowCount(_), type number}})
```

- The preceding formula is shown in the following screenshot:



Learning by example – Market Basket Analysis

- The formula must be replaced by the following:

```
= Table.Group(#"Changed Type", {"POS Txn"},
{{"Concat_dept", each Text.Combine([Dept]," | "), type
text}})
```

Learning by example – Market Basket Analysis

	A	B
1	POS Txn	Attributes
2	16120100160021008773	0261:HOSIERY | 0634:VITAMINS & HLTH AIDS | 0879:PET SUPPLIES | 0973:CANDY | 0982:SPIRITS | 0983:WINE | 0991:TOBACCO
3	16120100160021008774	0597:HEALTH AIDS | 0604:PERSONAL CARE
4	16120100160021008775	0819:PRE-RECORDED A/V | 0826:SMALL ELECTRICS | 0982:SPIRITS
5	16120100160021008776	0961:GENERAL GROCERIES
6	16120100160021008777	0982:SPIRITS
7	16120100160021008778	0982:SPIRITS | 0991:TOBACCO
8	16120100160021008779	0879:PET SUPPLIES | 0982:SPIRITS | 0983:WINE | 0984:BEER
9	16120100160021008780	0530:SCHOOL/OFFIC SUPP | 0597:HEALTH AIDS | 0601:VALUE ZONE | 0634:VITAMINS & HLTH AIDS | 0836:HOUSEHOLD CLEANING
10	16120100160021008781	0593:PRESTIGE COSMETICS | 0597:HEALTH AIDS | 0598:BABY CARE | 0836:HOUSEHOLD CLEANING | 0965:PERISHABLES | 0973:CANDY | 0983:WINE
11	16120100160021008782	0837:GENERAL HOUSEWARES | 0982:SPIRITS
12	16120100160021008783	0879:PET SUPPLIES | 0973:CANDY | 0984:BEER
13	16120100160021008784	0983:WINE
14	16120100160021008785	0962:BEVERAGES | 0982:SPIRITS
15	16120100160021008786	0982:SPIRITS | 0983:WINE
16	16120100160021008787	0982:SPIRITS
17	16120100160021008788	0638:GEN SPORTING GOODS | 0961:GENERAL GROCERIES | 0973:CANDY | 0991:TOBACCO
18	16120100160021008789	0646:SEASONAL | 0991:TOBACCO
19	16120100160021008790	0962:BEVERAGES | 0982:SPIRITS | 0983:WINE
20	16120100160021008791	0982:SPIRITS
21	16120100160021008792	0982:SPIRITS | 0983:WINE
22	16120100160021008793	0962:BEVERAGES | 0982:SPIRITS

Learning by example – Market Basket Analysis

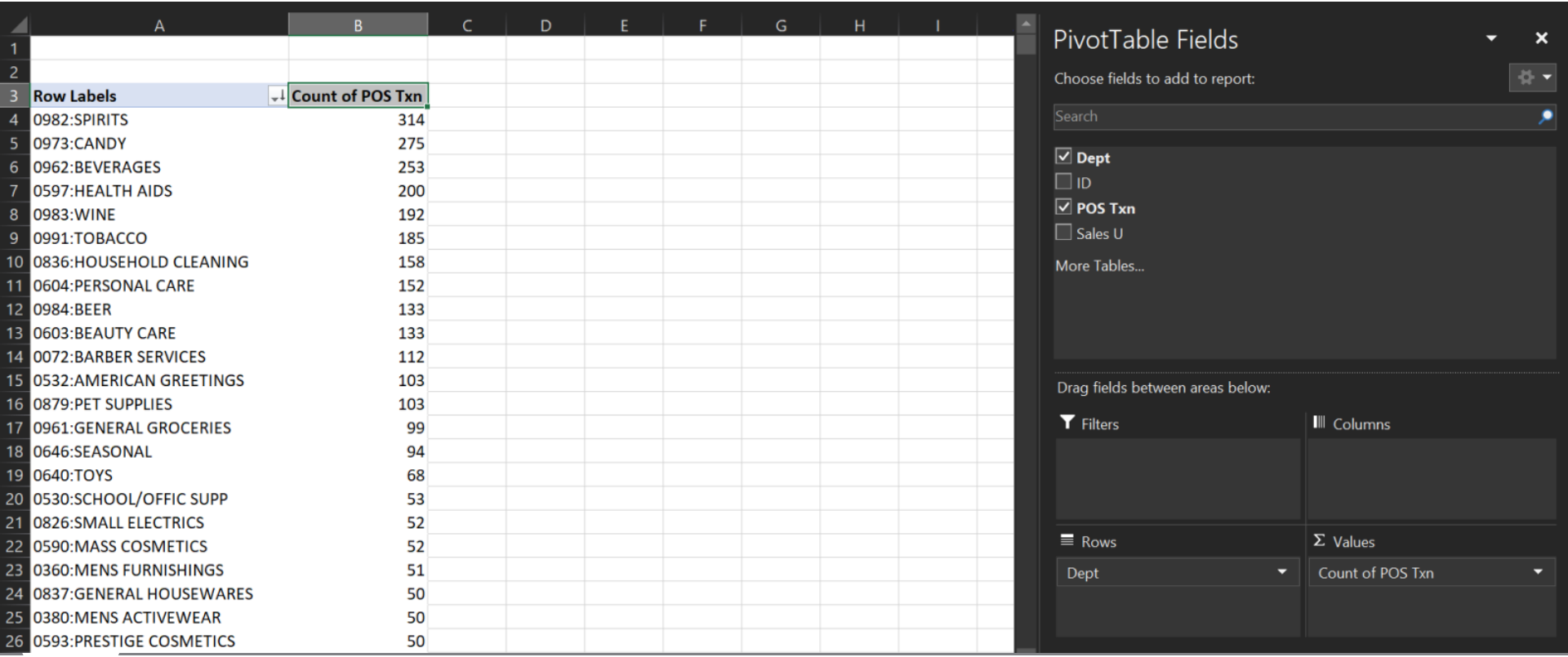
For each transaction ID, we now have a string representing the list of the departments involved. We are going to use this table to count the combinations of departments, but first, we convert it to a range:

- Right-click on any cell within the table and go to Table | Convert to Range.
- Rename the sheet to Concat depts (use the same name so that the references in future functions are correct).

- In a real-life example, we might have to limit the time period we are analyzing to reduce the amount of calculations needed and clean the data, leaving out unusual transactions (outliers).
- In our case, we will limit the number of combinations studied.
- We could take the departments in pairs, triads, or even larger numbers.
- The problem is that the number of combinations quickly scales with the number of departments. In fact, this number can be calculated as follows:

$$N_c = \binom{m}{n}$$

- Create a PivotTable as shown in the following screenshot:



Row Labels	Count of POS Txn
0982:SPIRITS	314
0973:CANDY	275
0962:BEVERAGES	253
0597:HEALTH AIDS	200
0983:WINE	192
0991:TOBACCO	185
0836:HOUSEHOLD CLEANING	158
0604:PERSONAL CARE	152
0984:BEER	133
0603:BEAUTY CARE	133
0072:BARBER SERVICES	112
0532:AMERICAN GREETINGS	103
0879:PET SUPPLIES	103
0961:GENERAL GROCERIES	99
0646:SEASONAL	94
0640:TOYS	68
0530:SCHOOL/OFFIC SUPP	53
0826:SMALL ELECTRICS	52
0590:MASS COSMETICS	52
0360:MENS FURNISHINGS	51
0837:GENERAL HOUSEWARES	50
0380:MENS ACTIVEWEAR	50
0593:PRESTIGE COSMETICS	50

Learning by example – Market Basket Analysis

- Order the rows by the count of transactions to get the top 10:

0982:SPIRITS
0973:CANDY
0962:BEVERAGES
0597:HEALTH AIDS
0983:WINE
0991:TOBACCO
0836:HOUSEHOLD CLEANING
0604:PERSONAL CARE
0603:BEAUTY CARE
0984:BEER

Learning by example – Market Basket Analysis

- Create two new columns that you can label X and Y.
- Define the first cell in column X as =$A$1 and the first cell in column Y as =A2.
- Select both cells and copy them down until column Y shows an empty value. Remember that the $ symbol fixes the cell value when copying.

Learning by example – Market Basket Analysis

- You will get a list as shown in the following screenshot, containing all possible X-Y pairs of departments:

C	D
X	Y
0982:SPIRITS	0973:CANDY
0982:SPIRITS	0962:BEVERAGES
0982:SPIRITS	0597:HEALTH AIDS
0982:SPIRITS	0983:WINE
0982:SPIRITS	0991:TOBACCO
0982:SPIRITS	0836:HOUSEHOLD CLEANING
0982:SPIRITS	0604:PERSONAL CARE
0982:SPIRITS	0603:BEAUTY CARE
0982:SPIRITS	0984:BEER

Learning by example – Market Basket Analysis

- The total number of elements should be as follows:

$$N_c = \binom{m}{n} = \binom{10}{2} = \frac{10!}{2!(10-2)!} = \frac{10 * 9}{2} = 45$$

Learning by example – Market Basket Analysis

- Here, we assume that the concatenated department names are in column B in the Concat depts sheet (this is why it needs to be named like that; if you understand the function you can change the name), and columns C and D contain the X and Y list respectively.
- The two COUNTIF functions account for the fact that the department names can appear in a different order, as shown in the following formula:

```
X&Y = =COUNTIF('Concat depts'!$B$2:$B$2065;"*"&C2&"*"&D2&"*")+COUNTIF('Concat depts'!$B$2:$B$2065;"*"&D2&"*"&C2&"*")
```

Learning by example – Market Basket Analysis

- We copy the following formulas until we reach the last element, that is, number 45:

X = COUNTIF('Concat depts'!$B$2:$B$2065;"*"&C2&"*")

- We will calculate the following:

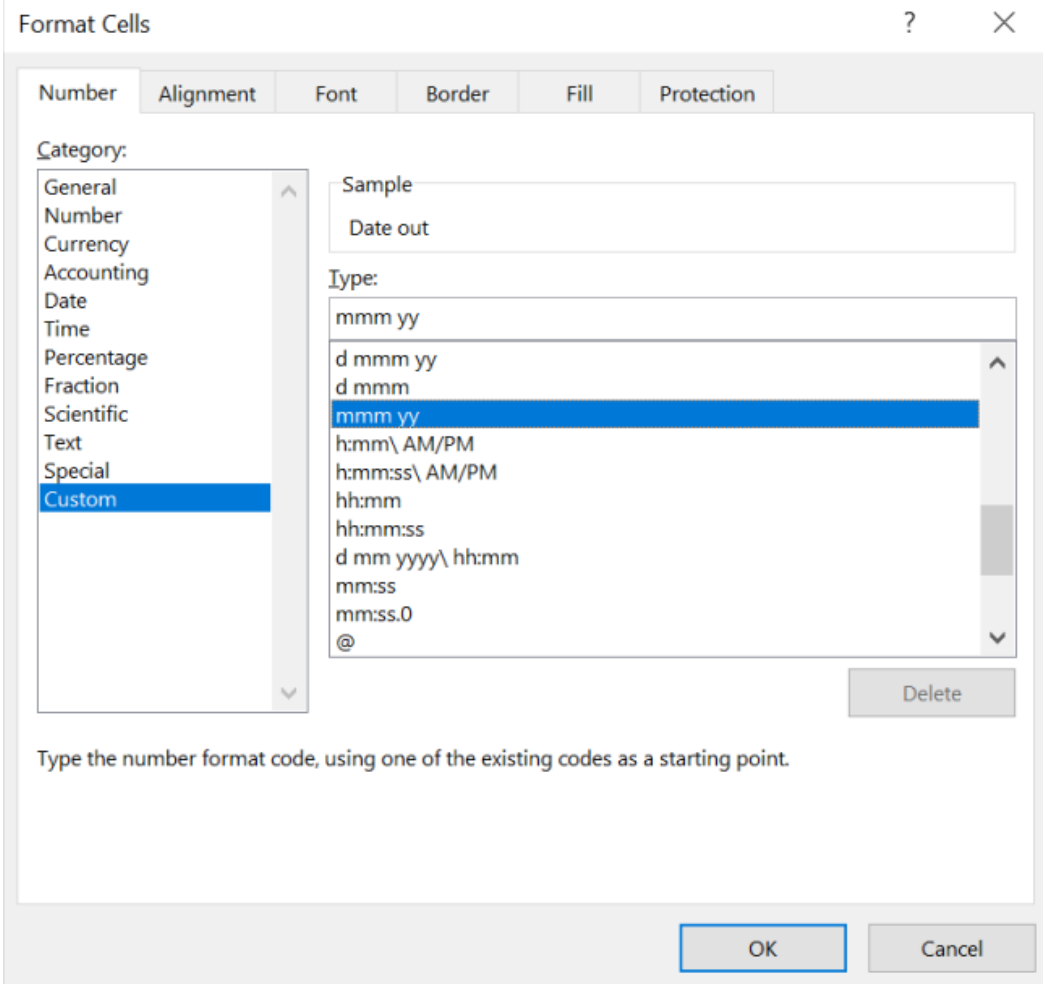
Support = X&Y/N

Confidence = X&Y/X

C	D	E	F	G	H	I
X	Y	X&Y	N	Support	X	Confidence
0982:SPIRITS	0973:CANDY	31	2064	2%	314	10%
0982:SPIRITS	0962:BEVERAGES	49	2064	2%	314	16%
0982:SPIRITS	0597:HEALTH AIDS	24	2064	1%	314	8%
0982:SPIRITS	0983:WINE	77	2064	4%	314	25%
0982:SPIRITS	0991:TOBACCO	52	2064	3%	314	17%
0982:SPIRITS	0836:HOUSEHOLD CLEANING	22	2064	1%	314	7%
0982:SPIRITS	0604:PERSONAL CARE	15	2064	1%	314	5%
0982:SPIRITS	0603:BEAUTY CARE	16	2064	1%	314	5%
0982:SPIRITS	0984:BEER	50	2064	2%	314	16%
0973:CANDY	0962:BEVERAGES	67	2064	3%	275	24%
0973:CANDY	0597:HEALTH AIDS	45	2064	2%	275	16%
0973:CANDY	0983:WINE	37	2064	2%	275	13%
0973:CANDY	0991:TOBACCO	25	2064	1%	275	9%
0973:CANDY	0836:HOUSEHOLD CLEANING	35	2064	2%	275	13%
0973:CANDY	0604:PERSONAL CARE	30	2064	1%	275	11%
0973:CANDY	0603:BEAUTY CARE	31	2064	2%	275	11%
0973:CANDY	0984:BEER	16	2064	1%	275	6%
0962:BEVERAGES	0597:HEALTH AIDS	36	2064	2%	253	14%
0962:BEVERAGES	0983:WINE	24	2064	1%	253	9%
0962:BEVERAGES	0991:TOBACCO	32	2064	2%	253	13%
0962:BEVERAGES	0836:HOUSEHOLD CLEANING	27	2064	1%	253	11%
0962:BEVERAGES	0604:PERSONAL CARE	30	2064	1%	253	12%
0962:BEVERAGES	0603:BEAUTY CARE	25	2064	1%	253	10%
0962:BEVERAGES	0984:BEER	18	2064	1%	253	7%
0597:HEALTH AIDS	0983:WINE	29	2064	1%	200	15%

Learning by example – Customer Cohort Analysis

- An excellent way of gaining insights about a company's customers and their behavior is to perform a segmented analysis.
- These segments are groups of customers that share the same characteristics and are usually called cohorts.
- Their definition depends very much on the type of business we are dealing with.


Format Cells
Number
Alignment
Font
Border
Fill
Protection
Category:
General
Number
Currency
Accounting
Date
Time
Percentage
Fraction
Scientific
Text
Special
Custom
Sample
Date out
Type:
mmm yy
d mmm yy
d mmm
mmm yy
h:mm\ AM/PM
h:mm:ss\ AM/PM
hh:mm
hh:mm:ss
d mm yyyy\ hh:mm
mm:ss
mm:ss.0
@
Delete
Type the number format code, using one of the existing codes as a starting point.
OK
Cancel

Learning by example – Customer Cohort Analysis

- Add two columns: Cohort, which in this case is equal to Date in, and Active months, which we can calculate as follows:

=DATEDIF(B2;C2;"m")

- This assumes that column B contains the Data in values and column C contains the Data out values.

	A	B	C	D	E	F
1	Id	Date in	Date out	Mean Monthly Spend	Cohort	Active months
2	236503	Feb-15	May-16	$ 203.90	Feb-15	15
3	236508	Feb-15	Aug-16	$ 547.80	Feb-15	23
4	236574	Feb-15	Jun-15	$ 865.80	Feb-15	22
5	236584	Feb-15	Aug-16	$ 408.20	Feb-15	16
6	236593	Feb-15	Nov-16	$ 455.40	Feb-15	21
7	236622	Feb-15	Sep-15	$ 387.60	Feb-15	14
8	236630	Feb-15	Jan-17	$ 156.90	Feb-15	23
9	236661	Feb-15	Oct-15	$ 941.20	Feb-15	7
10	236667	Feb-15	Apr-16	$ 195.80	Feb-15	7
11	236677	Feb-15	Jun-15	$ 869.60	Feb-15	9
12	236692	Feb-15	Jun-16	$ 692.10	Feb-15	6
13	236712	Feb-15	Jul-16	$ 878.30	Feb-15	11
14	236742	Feb-15	Jan-17	$ 918.80	Feb-15	20
15	236749	Feb-15	Apr-16	$ 452.30	Feb-15	8
16	236768	Feb-15	Nov-16	$ 181.10	Feb-15	3
17	236881	Feb-15	Jan-17	$ 121.40	Feb-15	5
18	236951	Feb-15	Jun-16	$ 279.90	Feb-15	23
19	236996	Feb-15	Aug-15	$ 334.80	Feb-15	14
20	237071	Feb-15	Sep-15	$ 774.40	Feb-15	11
21	237073	Feb-15	Jun-16	$ 304.20	Feb-15	16
22	237077	Feb-15	Aug-15	$ 646.30	Feb-15	3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Cohorts		Feb-15	Mar-15	Apr-15	May-15	Jun-15	Jul-15	Aug-15	Sep-15	Oct-15	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16
2	Month	0														
3	Month	1														
4	Month	2														
5	Month	3														
6	Month	4														
7	Month	5														
8	Month	6														
9	Month	7														
10	Month	8														
11	Month	9														
12	Month	10														
13	Month	11														
14	Month	12														
15	Month	13														
16	Month	14														
17	Month	15														
18	Month	16														
19	Month	17														
20	Month	18														
21	Month	19														
22	Month	20														

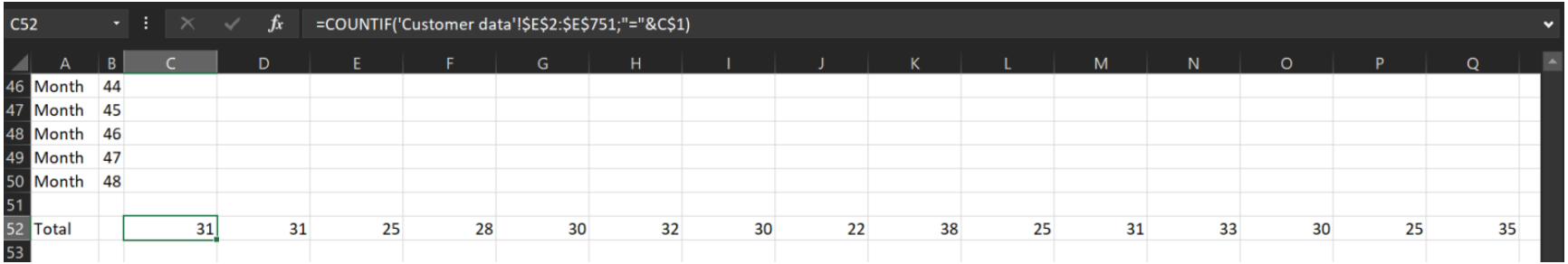
Learning by example – Customer Cohort Analysis

- Add a row at the end of the table to calculate the total number of customers in each cohort by using the following formula:

```
=COUNTIF('Customer data'!$E$2:$E$751;"="&C$1)
```

Learning by example – Customer Cohort Analysis

- Copy the cell contents to all cells to the right in the table, and, since the row is fixed, the correct values are calculated for each column, as shown in the following screenshot:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
46	Month	44															
47	Month	45															
48	Month	46															
49	Month	47															
50	Month	48															
51																	
52	Total		31	31	25	28	30	32	30	22	38	25	31	33	30	25	35
53																	

Learning by example – Customer Cohort Analysis

- In the first cell of the matrix, write the following formula:

```
=COUNTIFS('Customer data'!$E$2:$E$751;"="&C$1;'Customer data'!$C$2:$C$751;">"&EOMONTH(C$1;$B1))/C$52
```

Learning by example – Customer Cohort Analysis

- Copy the formula to the whole matrix.
- Format the cells as a percentage.
- Format the cells conditionally using a three-color scale. The result looks like the following screenshot in next slide

Learning by example – Customer Cohort Analysis

C2 =COUNTIFS('Customer data'!$E$2:$E$751;"="&C$1;'Customer data'!$C$2:$C$751;">"&EOMONTH(C$1;$B1))/C$52

Cohorts		Feb-15	Mar-15	Apr-15	May-15	Jun-15	Jul-15	Aug-15	Sep-15	Oct-15	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16	May-16	Jun-16
Month	0	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Month	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Month	2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Month	3	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Month	4	100.00%	96.77%	100.00%	92.86%	96.67%	96.88%	100.00%	90.91%	92.11%	100.00%	90.32%	87.88%	93.33%	96.00%	97.14%	96.43%	85.00%
Month	5	90.32%	93.55%	96.00%	89.29%	80.00%	90.63%	100.00%	90.91%	81.58%	100.00%	77.42%	84.85%	86.67%	92.00%	88.57%	92.86%	85.00%
Month	6	90.32%	90.32%	92.00%	78.57%	76.67%	87.50%	93.33%	90.91%	78.95%	92.00%	74.19%	81.82%	83.33%	92.00%	77.14%	82.14%	80.00%
Month	7	83.87%	83.87%	84.00%	78.57%	70.00%	84.38%	86.67%	90.91%	71.05%	88.00%	70.97%	75.76%	80.00%	88.00%	77.14%	78.57%	75.00%
Month	8	74.19%	80.65%	80.00%	78.57%	60.00%	81.25%	86.67%	90.91%	65.79%	80.00%	64.52%	72.73%	73.33%	88.00%	68.57%	75.00%	75.00%
Month	9	67.74%	74.19%	72.00%	71.43%	56.67%	75.00%	80.00%	77.27%	60.53%	76.00%	64.52%	72.73%	70.00%	80.00%	62.86%	64.29%	65.00%
Month	10	67.74%	70.97%	64.00%	67.86%	53.33%	71.88%	70.00%	72.73%	55.26%	68.00%	54.84%	69.70%	63.33%	68.00%	57.14%	64.29%	65.00%
Month	11	67.74%	67.74%	60.00%	64.29%	43.33%	65.63%	66.67%	68.18%	50.00%	64.00%	51.61%	69.70%	60.00%	52.00%	54.29%	60.71%	55.00%
Month	12	61.29%	61.29%	60.00%	57.14%	43.33%	65.63%	60.00%	68.18%	47.37%	60.00%	38.71%	60.61%	56.67%	52.00%	54.29%	57.14%	55.00%
Month	13	61.29%	51.61%	52.00%	50.00%	43.33%	59.38%	56.67%	59.09%	44.74%	52.00%	35.48%	54.55%	50.00%	48.00%	51.43%	50.00%	55.00%
Month	14	58.06%	48.39%	48.00%	46.43%	43.33%	56.25%	46.67%	45.45%	39.47%	52.00%	35.48%	54.55%	43.33%	44.00%	48.57%	46.43%	55.00%
Month	15	48.39%	45.16%	44.00%	39.29%	36.67%	53.13%	40.00%	40.91%	39.47%	44.00%	32.26%	45.45%	40.00%	44.00%	40.00%	46.43%	50.00%
Month	16	45.16%	45.16%	36.00%	35.71%	33.33%	50.00%	36.67%	36.36%	36.84%	40.00%	25.81%	39.39%	33.33%	32.00%	34.29%	46.43%	35.00%
Month	17	35.48%	45.16%	36.00%	28.57%	30.00%	43.75%	36.67%	22.73%	34.21%	36.00%	25.81%	33.33%	30.00%	32.00%	31.43%	42.86%	35.00%
Month	18	29.03%	45.16%	32.00%	21.43%	26.67%	40.63%	36.67%	22.73%	28.95%	28.00%	22.58%	30.30%	26.67%	28.00%	25.71%	35.71%	35.00%
Month	19	22.58%	41.94%	28.00%	14.29%	23.33%	34.38%	23.33%	13.64%	28.95%	24.00%	16.13%	24.24%	20.00%	28.00%	17.14%	32.14%	35.00%
Month	20	19.35%	32.26%	16.00%	10.71%	20.00%	34.38%	16.67%	13.64%	26.32%	20.00%	16.13%	21.21%	16.67%	24.00%	17.14%	32.14%	35.00%

Learning by example – Customer Cohort Analysis

We know now how many customers we were able to keep, but what was their value? They do not spend the same amount, so we need to include that variable in the analysis. Follow these steps:

- Create a similar matrix, with the month number and cohorts, but now we will use a different formula:

=SUMIFS('Customer data'!$D$2:$D$751;'Customer data'!$E$2:$E$751;"="&C$1;'Customer data'!$C$2:$C$751;">"&EOMONTH(C$1;$B1))

C2 =SUMIFS('Customer data'!$D$2:$D$751;'Customer data'!$E$2:$E$751;"="&C$1;'Customer data'!$C$2:$C$751;">"&EOMONTH(C$1;$B1))

Cohorts		Feb-15	Mar-15	Apr-15	May-15	Jun-15	Jul-15	Aug-15	Sep-15	Oct-15	Nov-15	Dec-15	Jan-16	Feb-16	Mar-16	Apr-16
Month	0	$ 16,532.50	$ 15,449.20	$ 12,329.30	$ 15,661.80	$ 15,636.40	$ 18,377.30	$ 15,741.20	$ 12,170.70	$ 23,526.00	$ 12,506.00	$ 16,640.30	$ 20,427.70	$ 16,234.40	$ 16,372.70	$ 20,611.20
Month	1	$ 16,532.50	$ 15,449.20	$ 12,329.30	$ 15,661.80	$ 15,636.40	$ 18,377.30	$ 15,741.20	$ 12,170.70	$ 23,526.00	$ 12,506.00	$ 16,640.30	$ 20,427.70	$ 16,234.40	$ 16,372.70	$ 20,611.20
Month	2	$ 16,532.50	$ 15,449.20	$ 12,329.30	$ 15,661.80	$ 15,636.40	$ 18,377.30	$ 15,741.20	$ 12,170.70	$ 23,526.00	$ 12,506.00	$ 16,640.30	$ 20,427.70	$ 16,234.40	$ 16,372.70	$ 20,611.20
Month	3	$ 16,532.50	$ 15,449.20	$ 12,329.30	$ 15,661.80	$ 15,636.40	$ 18,377.30	$ 15,741.20	$ 12,170.70	$ 23,526.00	$ 12,506.00	$ 16,640.30	$ 20,427.70	$ 16,234.40	$ 16,372.70	$ 20,611.20
Month	4	$ 16,532.50	$ 15,141.50	$ 12,329.30	$ 14,303.40	$ 15,251.30	$ 17,513.50	$ 15,741.20	$ 11,423.00	$ 22,145.60	$ 12,506.00	$ 14,828.30	$ 17,951.90	$ 15,095.30	$ 15,537.60	$ 20,448.80
Month	5	$ 13,906.20	$ 14,800.50	$ 11,539.90	$ 14,016.00	$ 12,388.50	$ 16,211.30	$ 15,741.20	$ 11,423.00	$ 19,253.10	$ 12,506.00	$ 12,197.10	$ 17,287.40	$ 14,155.70	$ 14,655.00	$ 18,635.00
Month	6	$ 13,906.20	$ 14,547.90	$ 11,324.70	$ 12,258.00	$ 11,636.40	$ 15,358.30	$ 14,443.30	$ 11,423.00	$ 18,620.90	$ 11,258.40	$ 11,849.80	$ 16,880.40	$ 13,506.00	$ 14,655.00	$ 16,609.50
Month	7	$ 12,925.10	$ 13,140.20	$ 10,741.40	$ 12,258.00	$ 10,891.80	$ 14,832.60	$ 13,439.40	$ 11,423.00	$ 17,542.30	$ 10,561.20	$ 11,214.20	$ 15,955.80	$ 13,158.60	$ 13,861.70	$ 16,609.50
Month	8	$ 11,316.80	$ 12,350.80	$ 9,872.80	$ 12,258.00	$ 9,568.50	$ 14,128.10	$ 13,439.40	$ 11,423.00	$ 15,692.00	$ 9,470.90	$ 10,339.20	$ 15,726.40	$ 11,471.80	$ 13,861.70	$ 15,275.40
Month	9	$ 9,637.10	$ 11,109.80	$ 9,287.60	$ 11,365.70	$ 8,672.80	$ 13,215.40	$ 12,881.50	$ 9,942.00	$ 14,364.50	$ 8,877.80	$ 10,339.20	$ 15,726.40	$ 10,474.40	$ 12,548.10	$ 13,973.30
Month	10	$ 9,637.10	$ 10,538.00	$ 8,551.70	$ 10,950.00	$ 8,530.50	$ 12,218.90	$ 11,263.80	$ 9,706.20	$ 12,764.50	$ 7,603.30	$ 8,677.70	$ 15,012.60	$ 9,579.80	$ 11,088.10	$ 13,011.20
Month	11	$ 9,637.10	$ 10,330.70	$ 7,732.60	$ 10,513.70	$ 6,277.00	$ 10,488.20	$ 10,304.20	$ 8,977.70	$ 11,824.70	$ 7,351.60	$ 8,084.60	$ 15,012.60	$ 9,166.50	$ 8,347.30	$ 12,469.40
Month	12	$ 8,353.70	$ 9,052.70	$ 7,732.60	$ 9,462.90	$ 6,277.00	$ 10,488.20	$ 9,375.40	$ 8,977.70	$ 11,224.70	$ 7,031.70	$ 5,090.80	$ 12,839.00	$ 8,492.40	$ 8,347.30	$ 12,469.40
Month	13	$ 8,353.70	$ 7,718.00	$ 6,625.40	$ 7,981.00	$ 6,277.00	$ 9,825.60	$ 9,194.60	$ 7,279.20	$ 10,791.30	$ 6,462.80	$ 4,152.20	$ 11,241.80	$ 7,329.00	$ 7,963.10	$ 12,242.00
Month	14	$ 7,996.60	$ 7,031.80	$ 6,011.70	$ 7,058.50	$ 6,277.00	$ 9,538.50	$ 7,653.80	$ 5,735.90	$ 9,802.40	$ 6,462.80	$ 4,152.20	$ 11,241.80	$ 6,920.30	$ 7,090.20	$ 11,523.40
Month	15	$ 7,035.50	$ 6,609.90	$ 5,049.10	$ 5,952.30	$ 5,878.80	$ 8,738.80	$ 7,340.50	$ 5,020.70	$ 9,802.40	$ 5,134.50	$ 3,954.30	$ 9,477.40	$ 6,217.80	$ 7,090.20	$ 9,350.90
Month	16	$ 6,831.60	$ 6,609.90	$ 3,720.70	$ 5,124.20	$ 5,739.00	$ 7,940.40	$ 6,575.50	$ 4,438.70	$ 9,443.40	$ 4,508.20	$ 2,958.60	$ 7,752.50	$ 5,540.10	$ 4,971.70	$ 7,514.70
Month	17	$ 5,555.40	$ 6,609.90	$ 3,720.70	$ 4,520.30	$ 4,798.90	$ 6,544.40	$ 6,575.50	$ 3,601.90	$ 8,741.60	$ 4,344.10	$ 2,958.60	$ 6,729.20	$ 5,335.50	$ 4,971.70	$ 7,080.40
Month	18	$ 3,824.40	$ 6,609.90	$ 3,443.80	$ 3,452.50	$ 4,300.70	$ 6,312.30	$ 6,575.50	$ 3,601.90	$ 7,564.50	$ 3,617.10	$ 1,959.50	$ 6,431.60	$ 4,885.20	$ 4,052.80	$ 6,281.70
Month	19	$ 2,868.40	$ 6,476.90	$ 3,326.60	$ 2,111.00	$ 3,558.10	$ 5,186.10	$ 3,468.20	$ 2,343.20	$ 7,564.50	$ 3,181.00	$ 1,431.50	$ 5,537.40	$ 3,510.50	$ 4,052.80	$ 3,501.90
Month	20	$ 2,373.40	$ 5,175.10	$ 1,686.70	$ 1,633.00	$ 3,367.00	$ 5,186.10	$ 1,999.10	$ 2,343.20	$ 6,750.50	$ 2,872.90	$ 1,431.50	$ 4,717.20	$ 2,667.10	$ 3,262.60	$ 3,501.90

Summary

- In this lesson, we learned about two data mining techniques: Market Basket Analysis and Customer Cohort Analysis.
- The first one tells us about hidden relations between store departments or products based on the customers' behavior.
- The second shows the time evolution of the number of customers, revealing differences between different customer segments or cohorts.

7 : Implementing Time Series

LEARNING VOYAGE

Implementing Time Series

In this lesson, we will cover the following topics:

- Modeling and visualizing time series
- Forecasting time series automatically in Excel
- Studying the stationarity of a time series

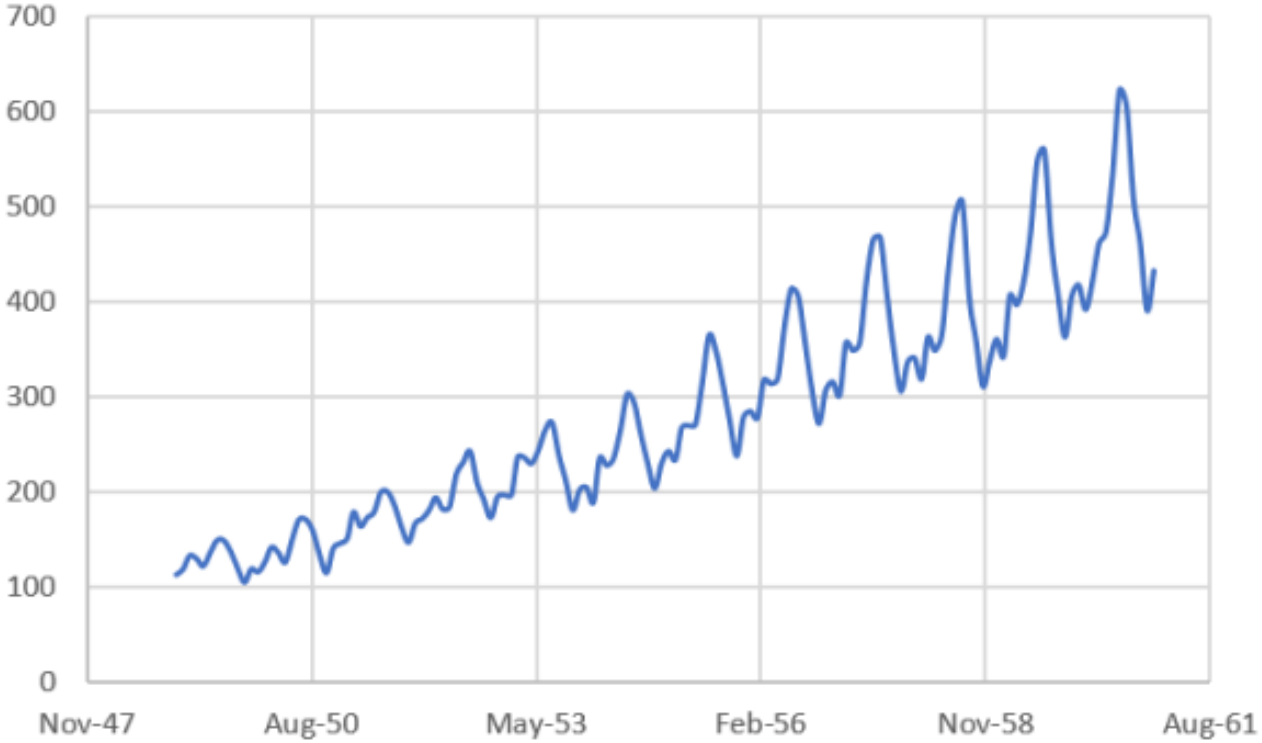
Technical requirements

- To complete this lesson, the reader will need to download the AirPassengers_modified.csv file from the GitHub repository

Modeling and visualizing time series

- We have seen that doing a preliminary data analysis and visualizing the dataset is the first step in any machine learning project.
- Time series are no exception. So, we will start by exploring time series and learning about its different characteristics.
- In the case of a time series, a preliminary analysis implies modeling it; that is, understanding whether it is periodic, whether it shows a given tendency (increasing or decreasing with time), or whether it is stationary (mean and variance of the values don't change over time), among other measures.

Modeling and visualizing time series



Modeling and visualizing time series

The easiest way to get a trend is to use Excel's built-in capability to calculate trends. To use it, follow these steps:

- Click on the chart area.
- Check the box for Trendline

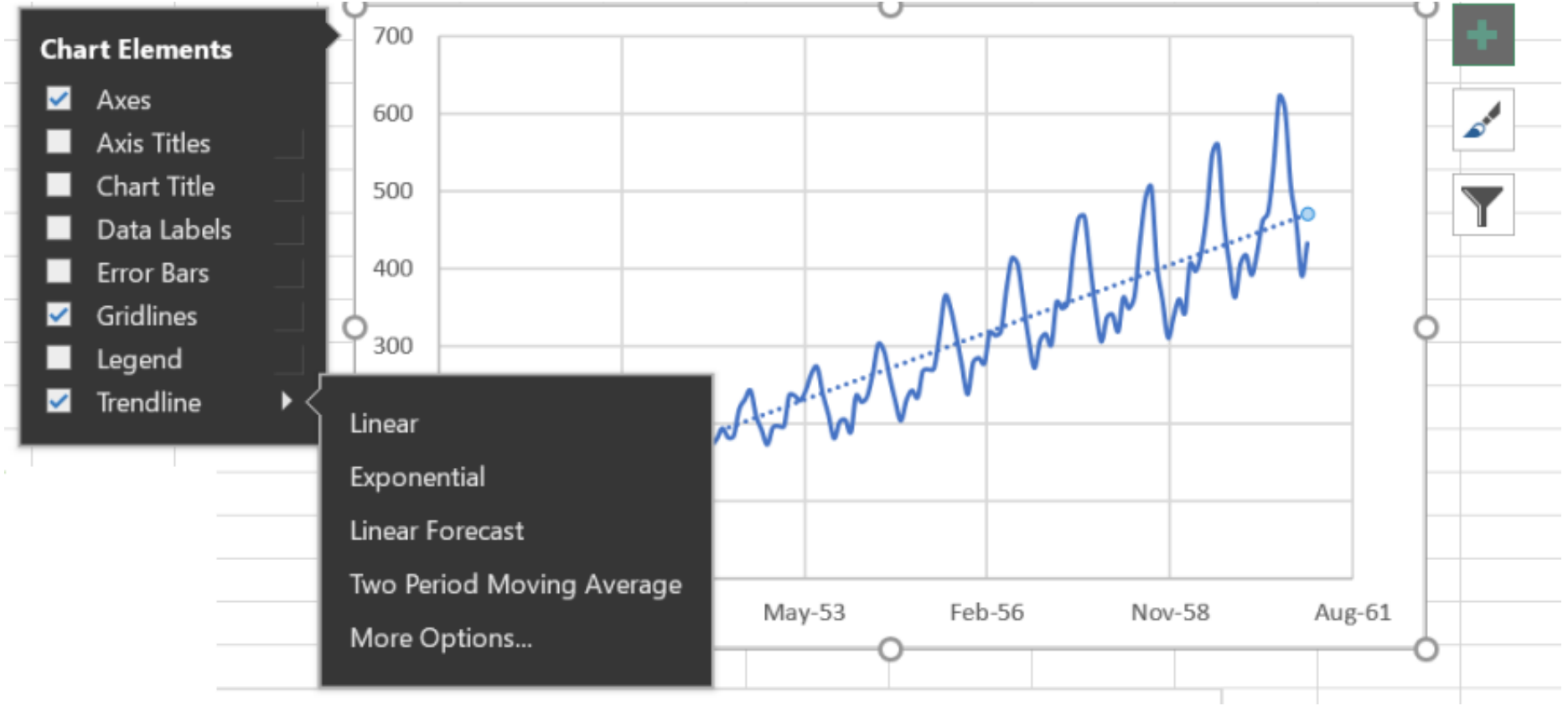
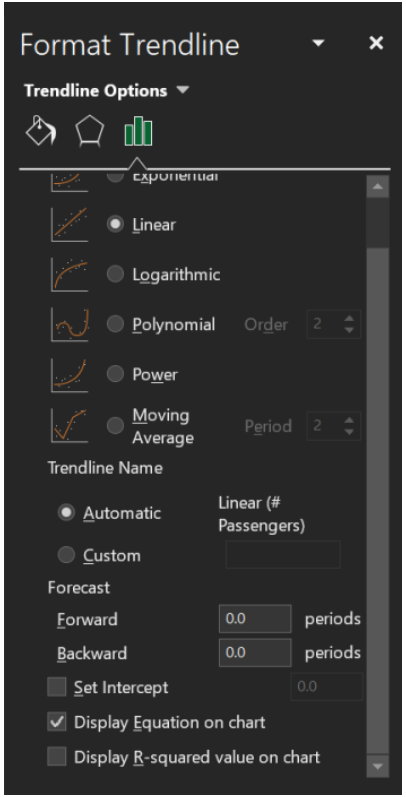
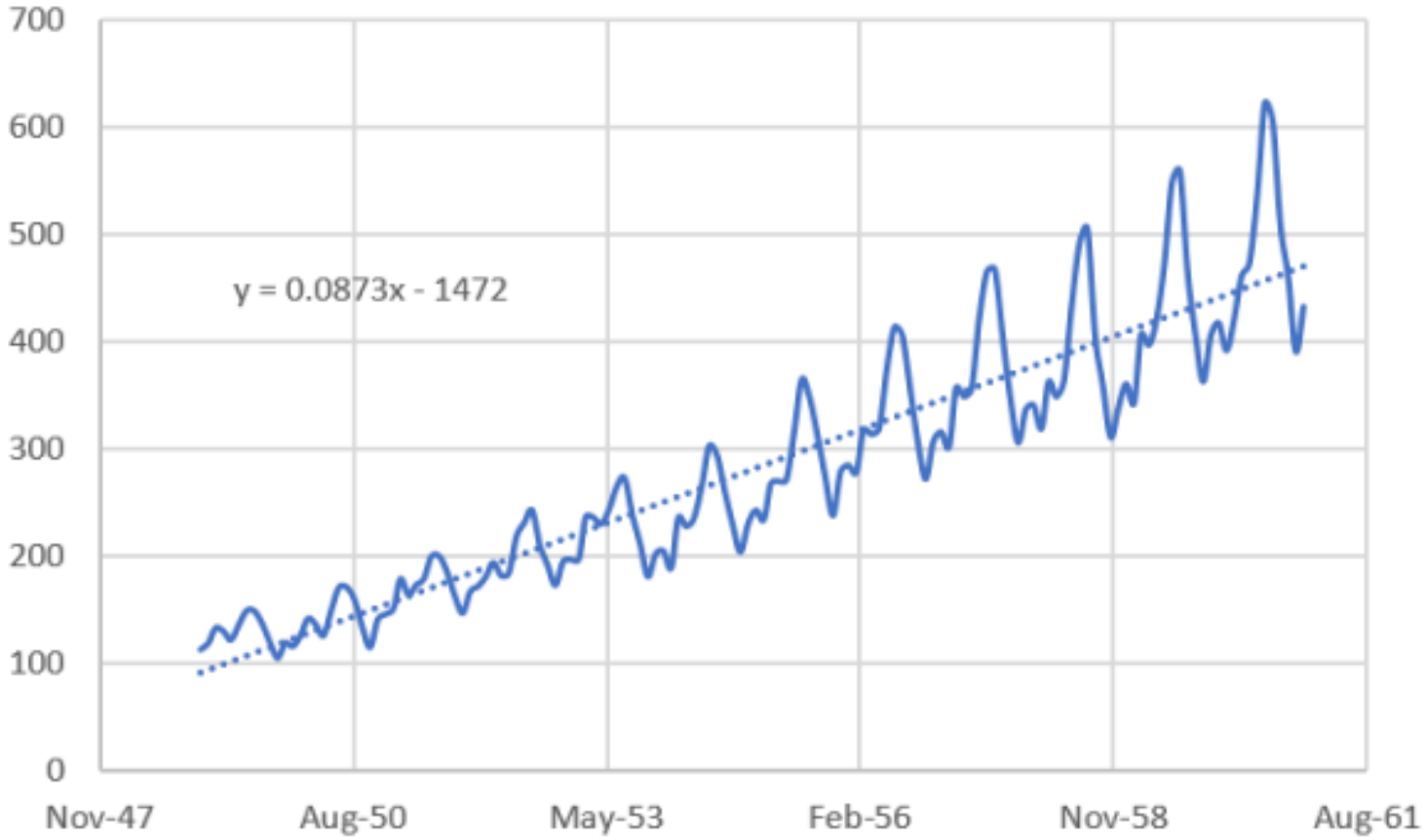


Chart Elements
Axes
Axis Titles
Chart Title
Data Labels
Error Bars
Gridlines
Legend
Trendline
Linear
Exponential
Linear Forecast
Two Period Moving Average
More Options...
700
600
500
400
300
May-53
Feb-56
Nov-58
Aug-61

Modeling and visualizing time series

- Notice that Linear is not the only option for a trend line and that more sophisticated regressions are available, as shown in the following screenshot:




700
600
500
400
300
200
100
0
y = 0.0873x - 1472
Nov-47
Aug-50
May-53
Feb-56
Nov-58
Aug-61

Modeling and visualizing time series

- A periodic part, which repeats every 12 months (we guess this because of the periodicity of the peaks, and call this 12 month period a season)
- An increasing part, which we can obtain by regression or averaging the series
- A noise part, which we basically define as the remaining values once we isolate the first two parts

Modeling and visualizing time series

- This model can then be written as follows:

$$Passengers = periodic(TravelDate) * increasing(TravelDate) * noise(TravelDate)$$

Modeling and visualizing time series

- Start by calculating the average number of passengers in the first 12 months:

=AVERAGE(B2:B13)

- In the cell to the right, calculate the standard deviation (we will use it in the next section, when we test the stationarity of the series):

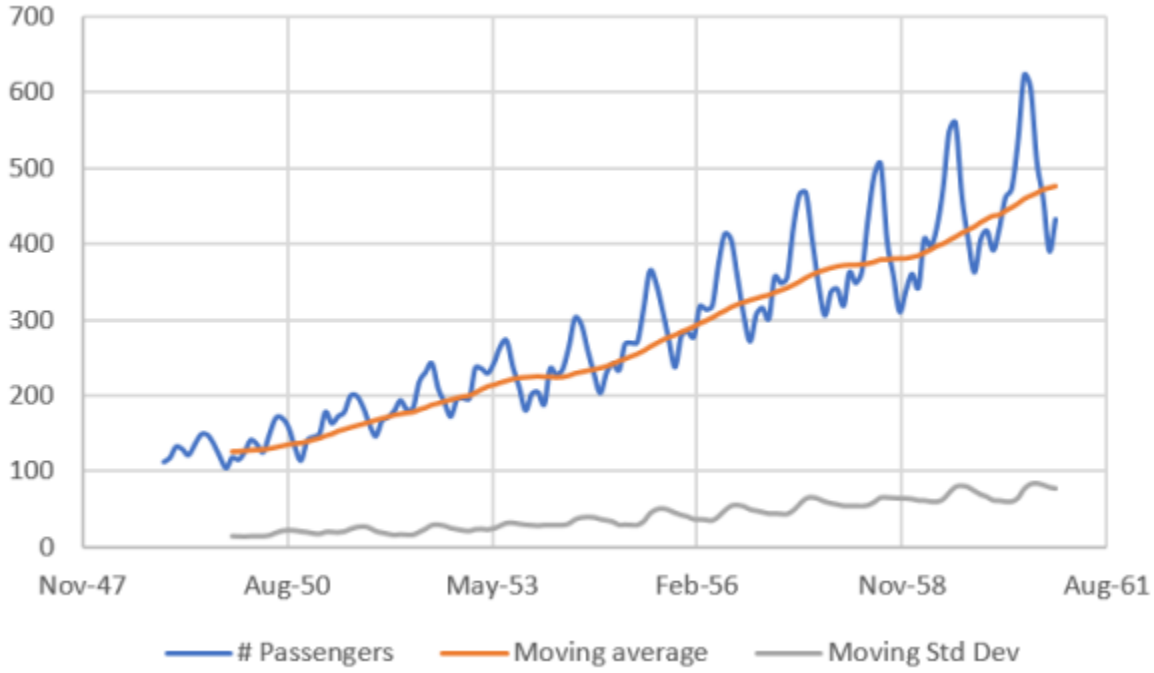
=STDEV.S(B2:B13)

Modeling and visualizing time series

- Copy both calculations down to the end of the table.

TravelDate	# Passengers	Moving average	Moving Std Dev
Jan-49	112		
Feb-49	118		
Mar-49	132		
Apr-49	129		
May-49	121		
Jun-49	135		
Jul-49	148		
Aug-49	148		
Sep-49	136		
Oct-49	119		
Nov-49	104		
Dec-49	118	126.6666667	13.72014666
Jan-50	115	126.9166667	13.45334249
Feb-50	126	127.5833333	13.16647487
Mar-50	141	128.3333333	13.68697678
Apr-50	135	128.8333333	13.82246744
May-50	125	129.1666667	13.66370995
Jun-50	149	130.3333333	14.76071773

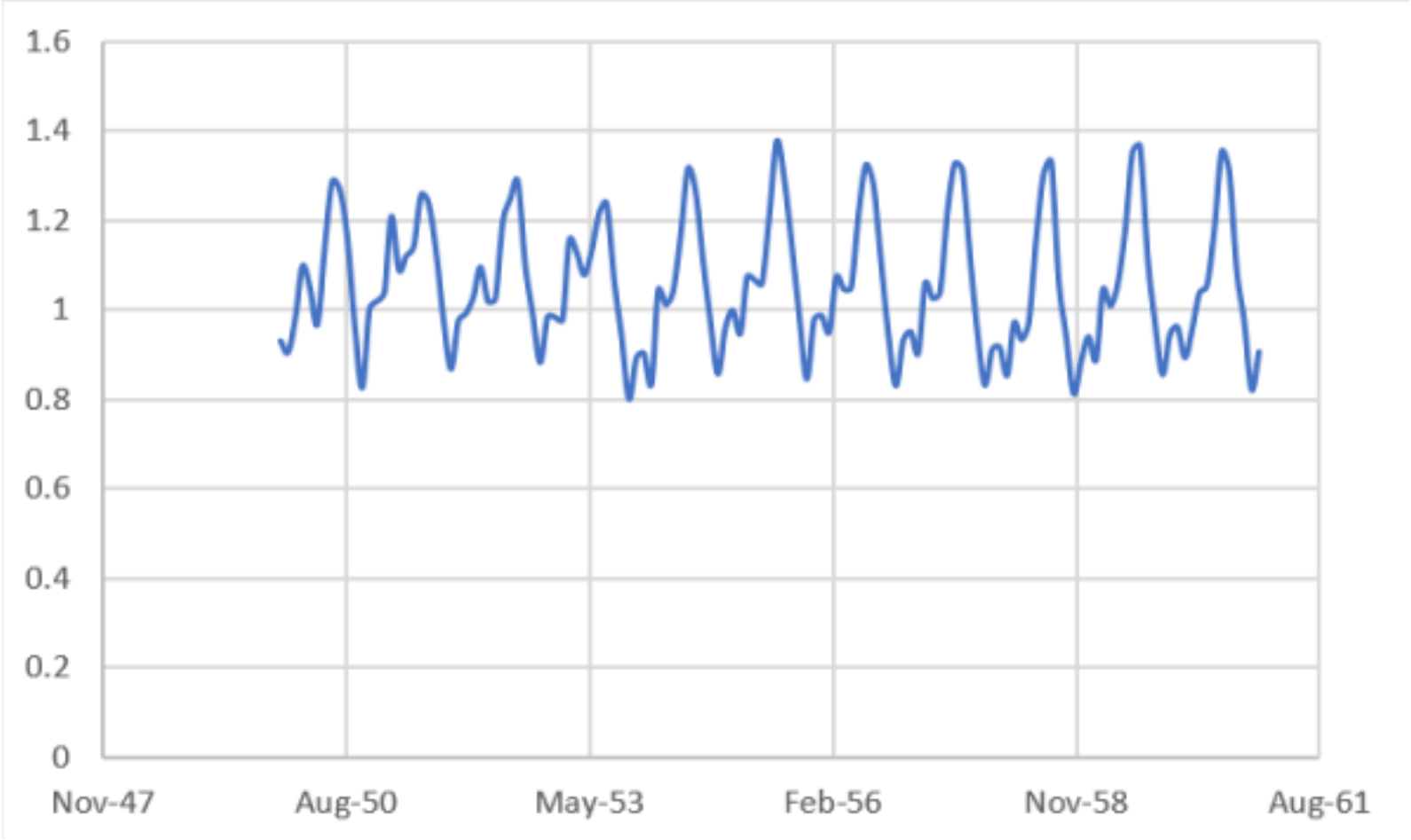
Modeling and visualizing time series



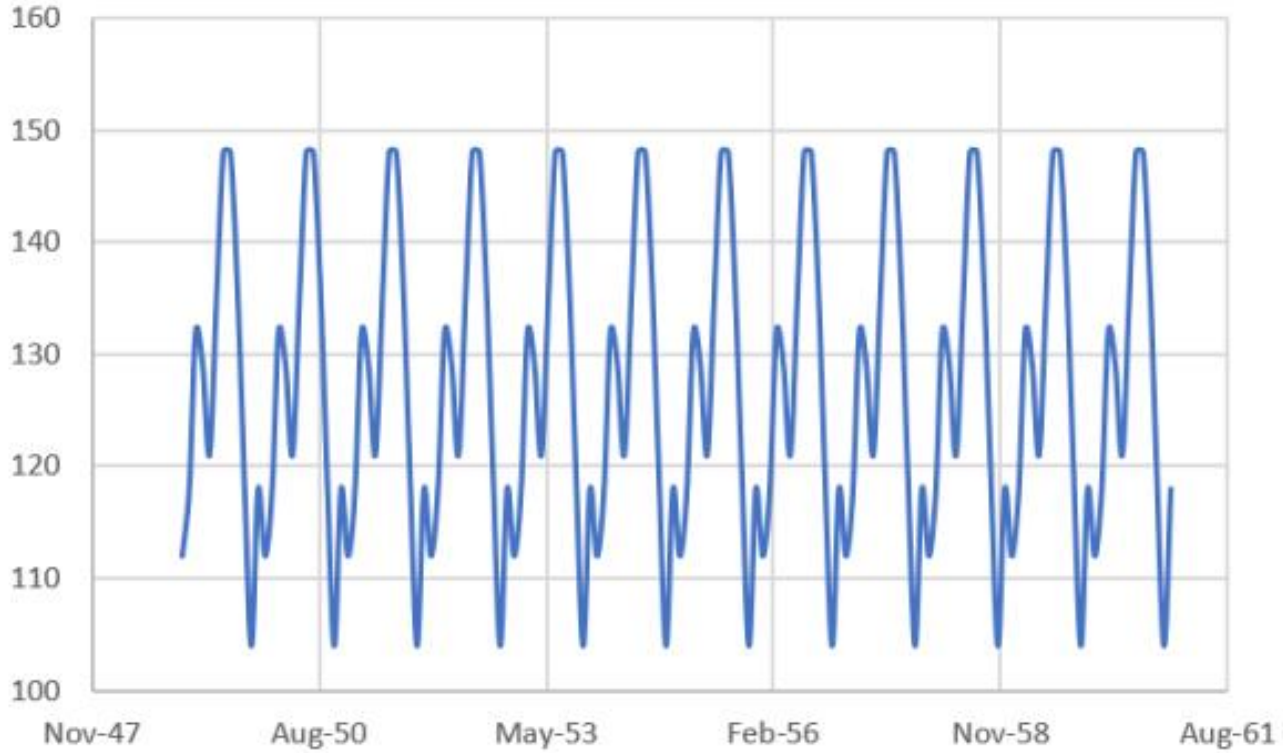
Modeling and visualizing time series

- Going back to our model for the time series, we can write:

$$\frac{Passengers}{increasing(TravelDate)} = periodic(TravelDate) * noise(TravelDate)$$


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Nov-47
Aug-50
May-53
Feb-56
Nov-58
Aug-61

Modeling and visualizing time series

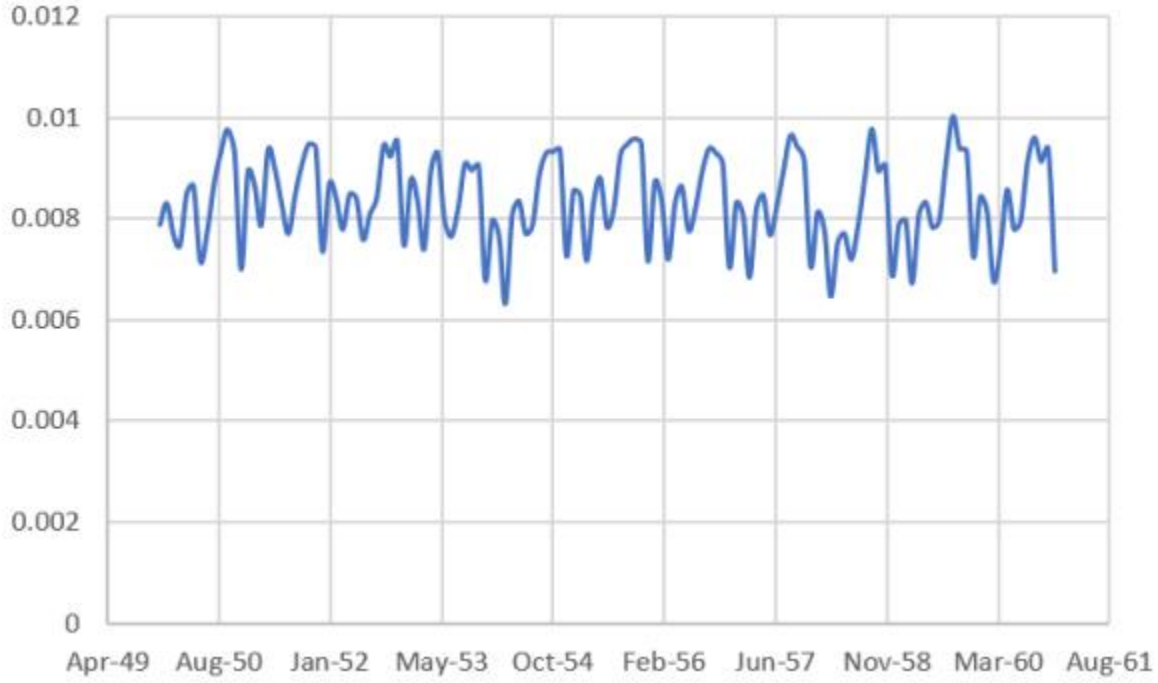


Modeling and visualizing time series

- Finally, we will calculate noise(TravelDate). We will, again, calculate this from our model:

$$noise(TravelDate) = \frac{Passengers}{periodic(TravelDate) * increasing(TravelDate)}$$

- In another column, use the preceding calculation to create a new diagram:



Modeling and visualizing time series

- Open a new sheet.
- Copy the series values and extend the time period up to Dec-62.
- In column C, copy the periodic(TravelDate) values (24 values in two equal series of 12 values).
- In column D, copy the noise values corresponding to the last two years.
- In column E, cell 146, calculate increasing(TravelDate) using the trendline formula (=0.0873*A146 - 1472).

Modeling and visualizing time series

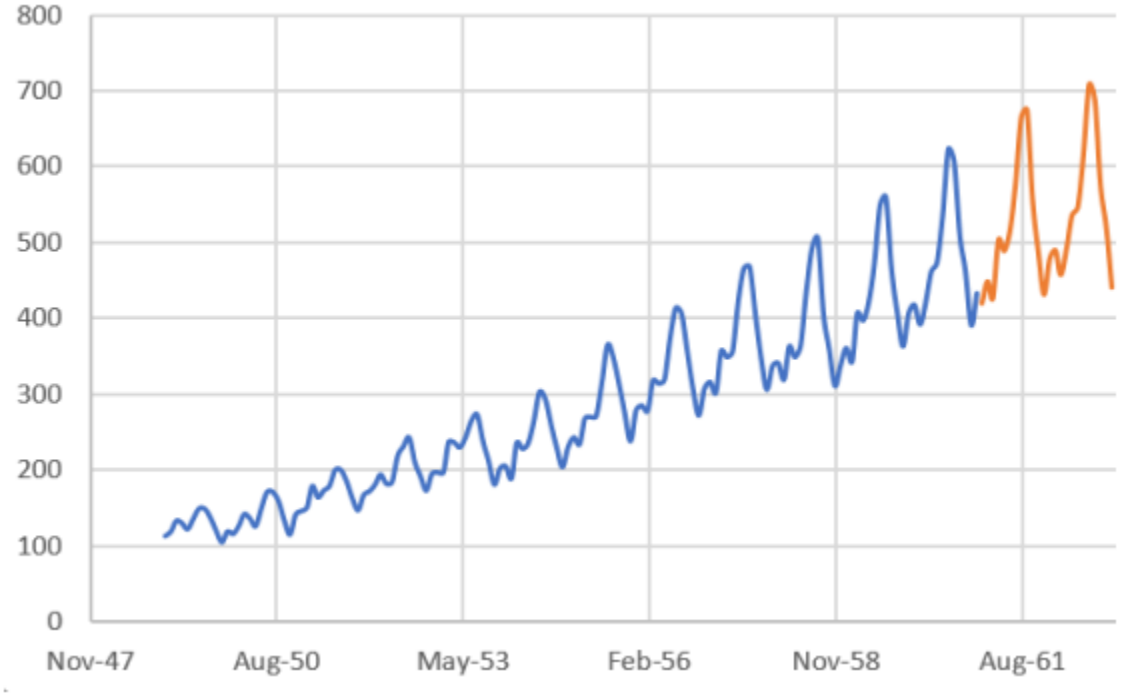
- Copy the formula down to the end of the table. Cell B1 is then as follows:

=C146*D146*E146

- The same calculation is then copied down.
- The resulting table is as follows:

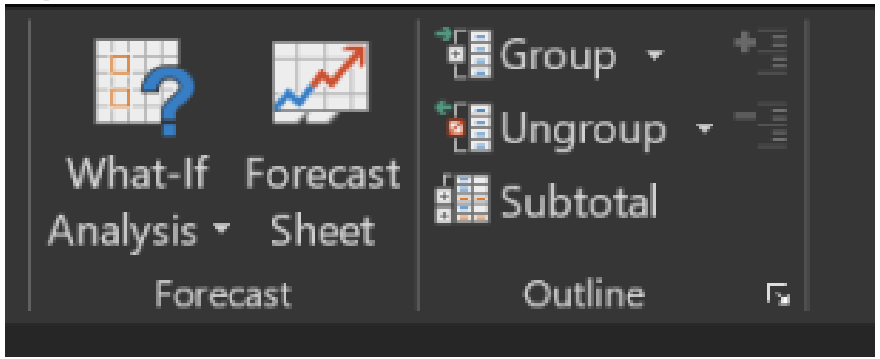
	A	B	C	D	E
144	Nov-60	390			
145	Dec-60	432			
146	Jan-61	418.5687	112	0.007897	473.22
147	Feb-61	447.7342	118	0.007973	475.92
148	Mar-61	425.3093	132	0.006735	478.37
149	Apr-61	502.9614	129	0.008105	481.08
150	May-61	488.2151	121	0.008342	483.69
151	Jun-61	514.4723	135	0.007835	486.40
152	Jul-61	576.8031	148	0.00797	489.02
153	Aug-61	665.485	148	0.009144	491.73
154	Sep-61	675.077	136	0.010039	494.43
155	Oct-61	555.4341	119	0.00939	497.05
156	Nov-61	486.2185	104	0.009355	499.76
157	Dec-61	430.2691	118	0.007258	502.38
158	Jan-62	477.5688	112	0.008442	505.08
159	Feb-62	488.9317	118	0.00816	507.79
160	Mar-62	456.3509	132	0.006776	510.23
161	Apr-62	490.4094	129	0.007411	512.94
162	May-62	535.7012	121	0.008587	515.56
163	Jun-62	546.0296	135	0.007804	518.27
164	Jul-62	614.8332	148	0.007975	520.88
165	Aug-62	708.8846	148	0.009148	523.59
166	Sep-62	688.351	136	0.009617	526.30
167	Oct-62	575.2492	119	0.009139	528.92
168	Nov-62	519.6916	104	0.0094	531.62
169	Dec-62	439.6429	118	0.006974	534.24

Modeling and visualizing time series



Forecasting time series automatically in Excel

- Select both columns, TravelDate and Passengers, corresponding to the time and number of passengers.
- Navigate to Data in the main menu.
- Select Forecast Sheet (see the following screenshot for reference):



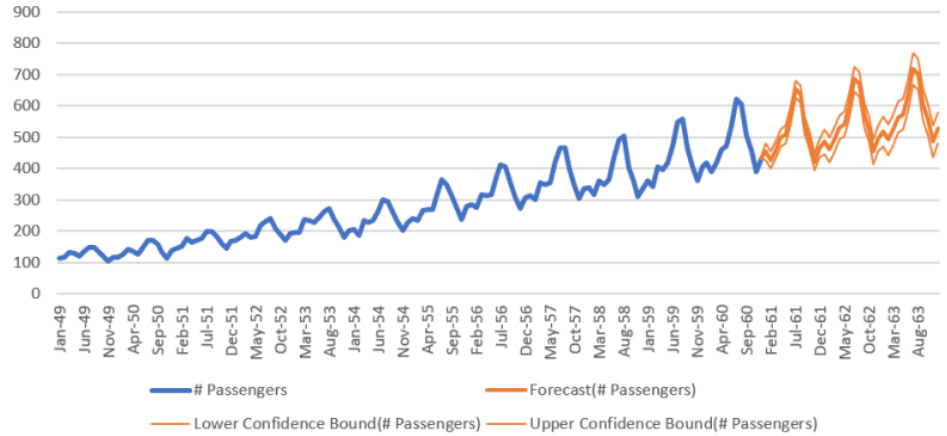
Forecasting time series automatically in Excel

A window will pop up, showing a preview of the forecast and giving us the chance to change some parameters by clicking in Options:

- Forecast End
- Forecast Start
- Confidence interval
- Seasonality

Create Forecast Worksheet

Use historical data to create a visual forecast worksheet



Forecast End: 1 12 1963

Options

Forecast Start: 1 12 1960

Confidence Interval: 95%

Seasonality

- Detect Automatically
- Set Manually: 12

Include forecast statistics

Timeline Range: Sheet6!$A$1:$A$145

Values Range: Sheet6!$B$1:$B$145

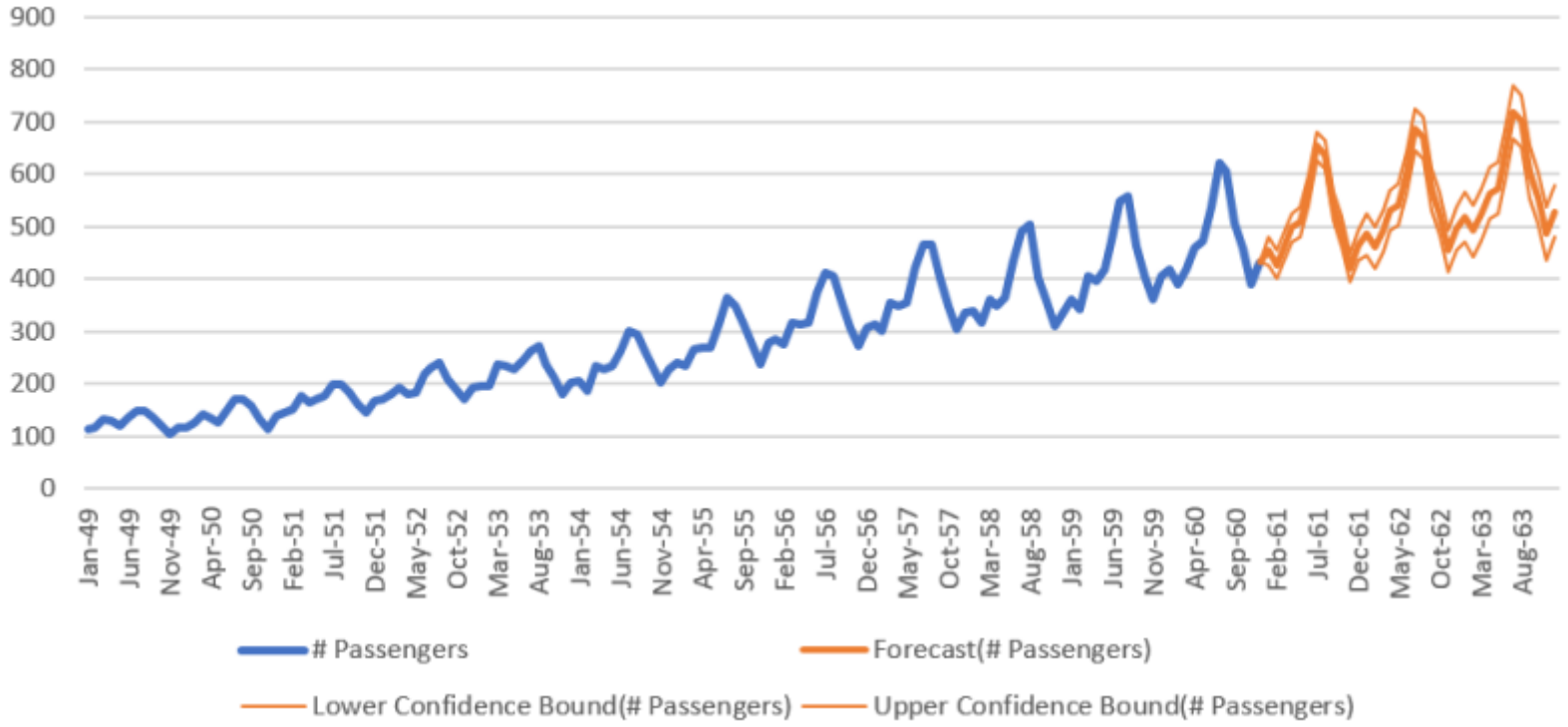
Fill Missing Points Using: Interpolation

Aggregate Duplicates Using: Average

Create | Cancel

	TravelDate	# Passengers	Forecast(# Passengers)	Lower Confidence Bound(# Passengers)	Upper Confidence Bound(# Passengers)
142	Sep-60	508			
143	Oct-60	461			
144	Nov-60	390			
145	Dec-60	432	432	432.00	432.00
146	Jan-61		453.8122736	427.08	480.54
147	Feb-61		427.8263121	400.96	454.69
148	Mar-61		459.262672	432.26	486.27
149	Apr-61		498.909199	471.76	526.05
150	May-61		508.8944168	481.61	536.18
151	Jun-61		569.8482444	542.42	597.28
152	Jul-61		653.6777865	626.10	681.25
153	Aug-61		637.3598214	609.63	665.08
154	Sep-61		539.05814	511.18	566.93
155	Oct-61		490.8222267	462.80	518.85
156	Nov-61		421.258682	393.08	449.44
157	Dec-61		464.425986	436.09	492.76
158	Jan-62		486.2264429	447.07	525.38
159	Feb-62		460.2404815	420.97	499.51
160	Mar-62		491.6768414	452.28	531.07
161	Apr-62		531.3233684	491.81	570.84
162	May-62		541.3085862	501.68	580.94
163	Jun-62		602.2624138	562.51	642.02

Forecasting time series automatically in Excel



Forecasting time series automatically in Excel

- If we select Include forecast statistics in the pop-up window, we get the following table:

Statistic	Value
Alpha	0.10
Beta	0.00
Gamma	0.90
MASE	0.69
SMAPE	0.03
MAE	14.58
RMSE	17.07

Studying the stationarity of a time series

There are three main checks of stationarity in practice:

- The mean value is constant (does not depend on time).
- The variance is constant.
- The covariance of the elements i and i+m is constant.

Studying the stationarity of a time series

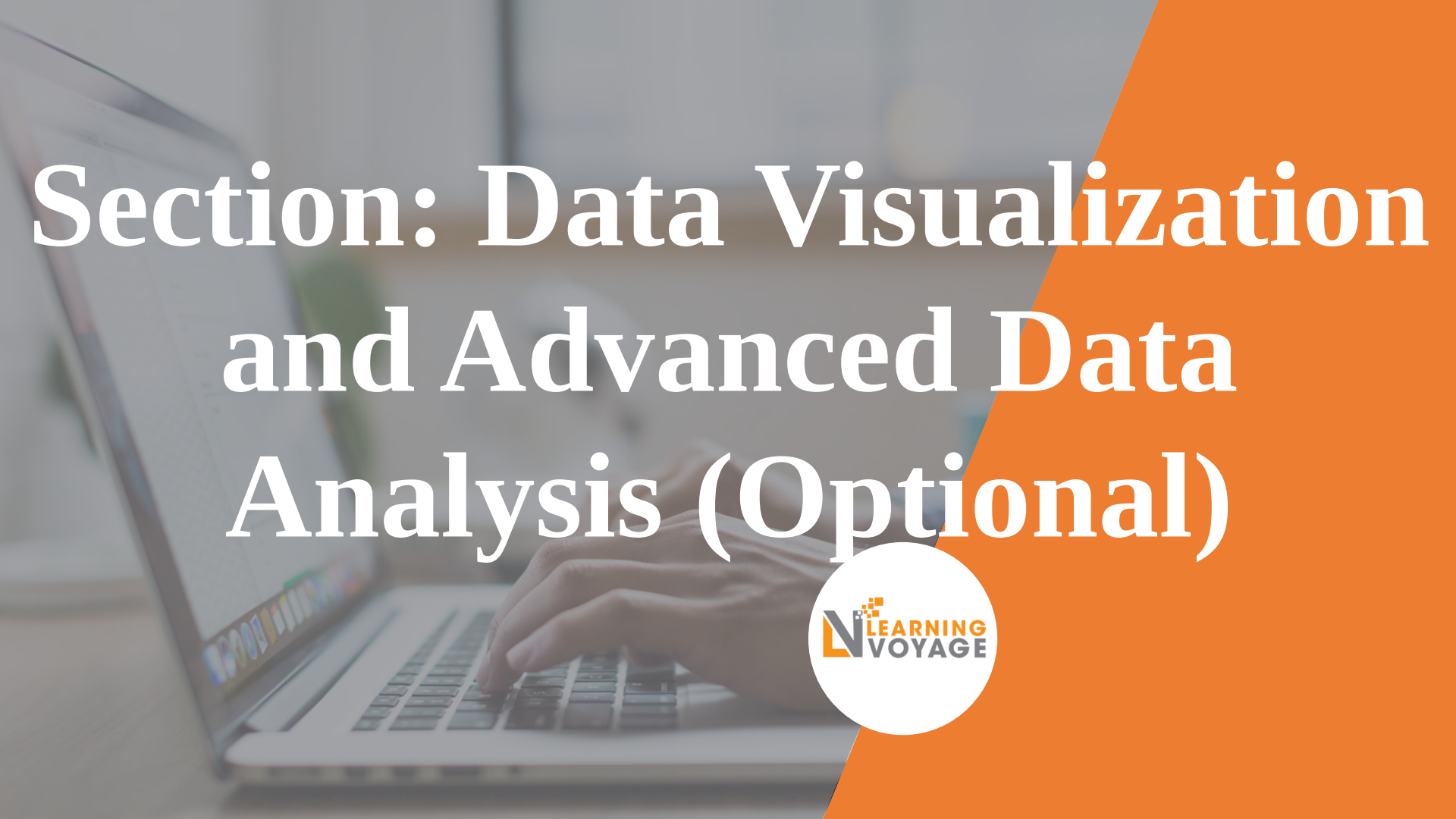
There are two ways of removing seasonality and trend:

- The technique of decomposing the series into noise, periodic, and increasing terms.

- Differencing – that is, creating a new series by taking the difference between the values i and (i+m). The position difference, m is known as lag.

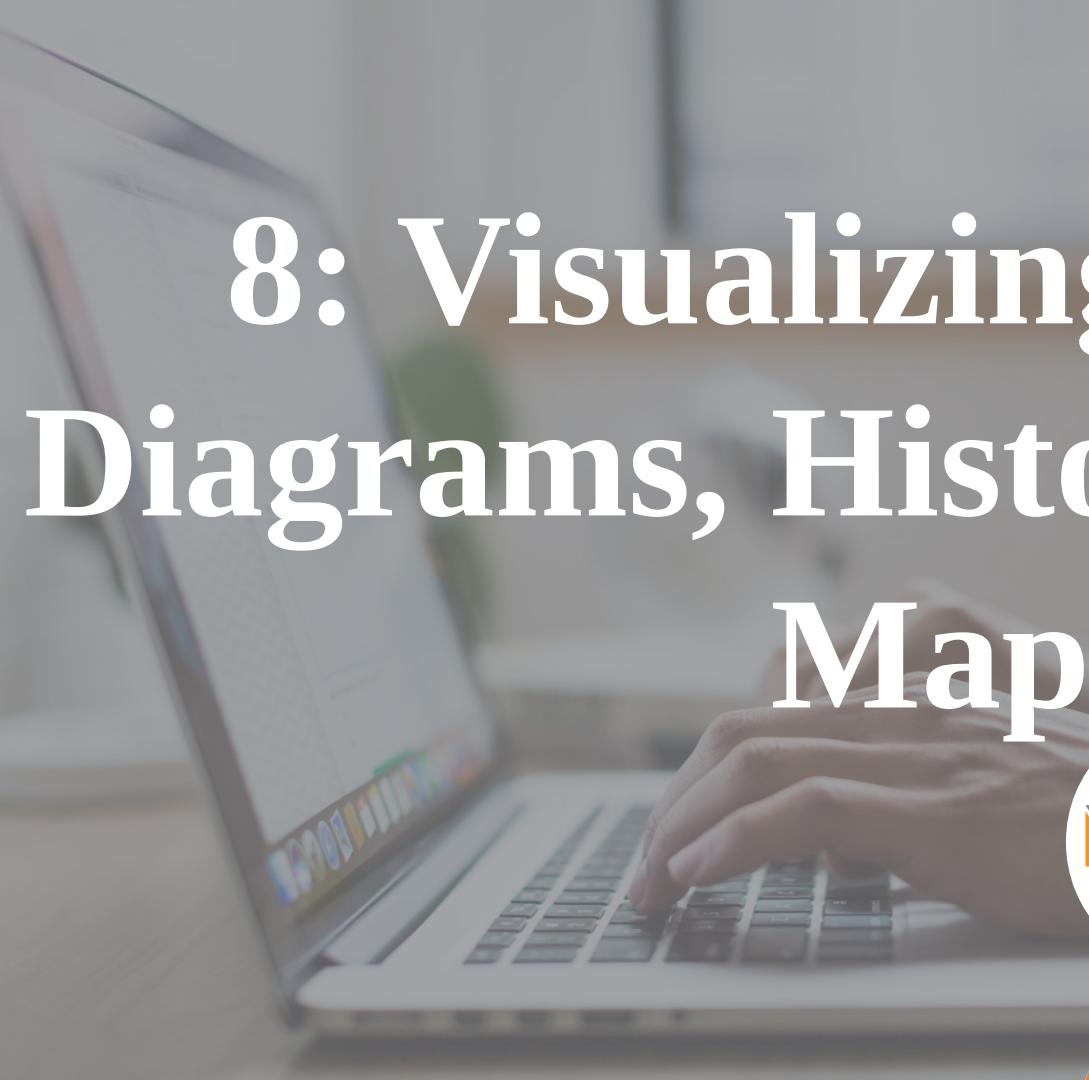
Summary

- We have seen a step-by-step method to decompose a time series and forecast its future values.
- This can help us, at least in general terms, to predict the outcome of different processes.
- Time series can be studied both graphically and numerically, extracting their characteristics and using them to understand how they will behave in the future.



Section: Data Visualization and Advanced Data Analysis (Optional)

LEARNING VOYAGE



8: Visualizing Data in Diagrams, Histograms, and Maps

LEARNING VOYAGE

Visualizing Data in Diagrams, Histograms, and Maps

The following topics will be covered in this lesson:

- Showing basic comparisons and relationships between variables
- Building data distributions using histograms
- Representing geographical distribution of data in maps
- Showing data that changes over time

Technical requirements

- To complete this lesson, the reader will need to download the 1976USpresident.xlsx and subte.xlsx files from the GitHub repository

Showing basic comparisons & relationships b/w variables

- Tell the story of your data and help decision makers with their job.
- Predict the future evolution of some variable(s).
- Find hidden trends and patterns in the data.
- Find outliers, that is, anomalies in the data.
- Understand the distribution, composition, and relationships.
- Build groups and categories.

Showing basic comparisons & relationships b/w variables

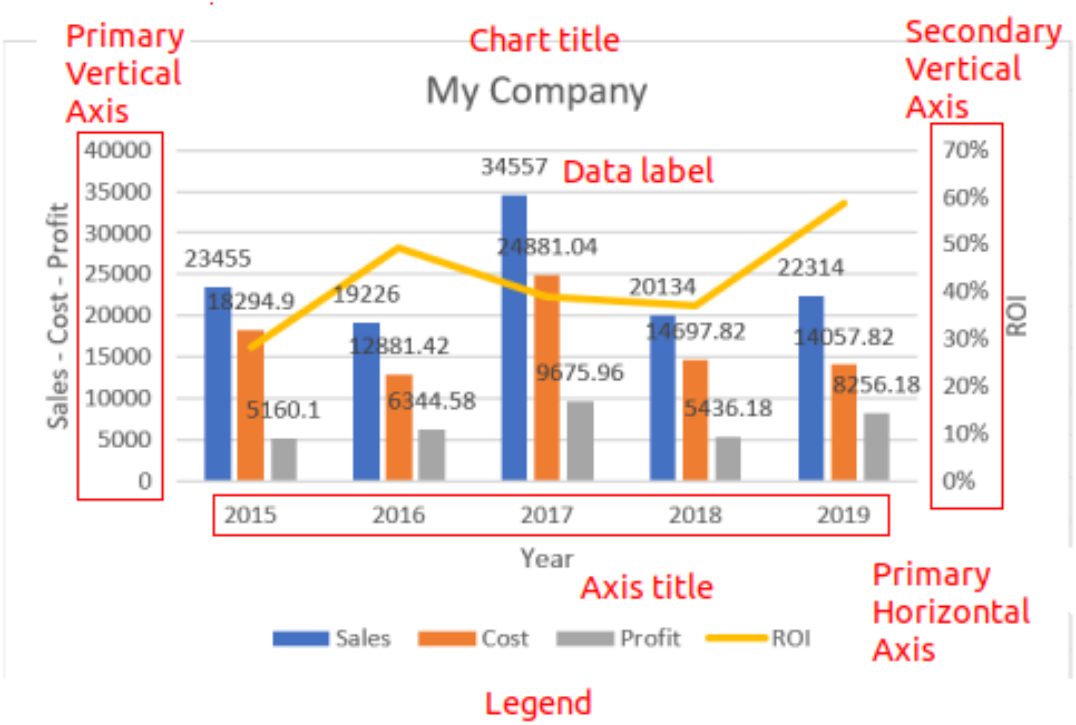
Year	Sales	Cost	Profit	ROI
2015	23455	18294.9	5160.1	28.21%
2016	19226	12881.42	6344.58	49.25%
2017	34557	24881.04	9675.96	38.89%
2018	20134	14697.82	5436.18	36.99%
2019	22314	14057.82	8256.18	58.73%

Showing basic comparisons & relationships b/w variables

Year	SalesA	CostA	ProfitA	SalesB	CostB	ProfitB
2015	23455	9	1	23455	9	1
2016	19226	42	58	19226	42	58
2017	34557	04	96	34557	04	96
2018	20134	82	18	82	18	
2019	22314	82	18	22314	82	18

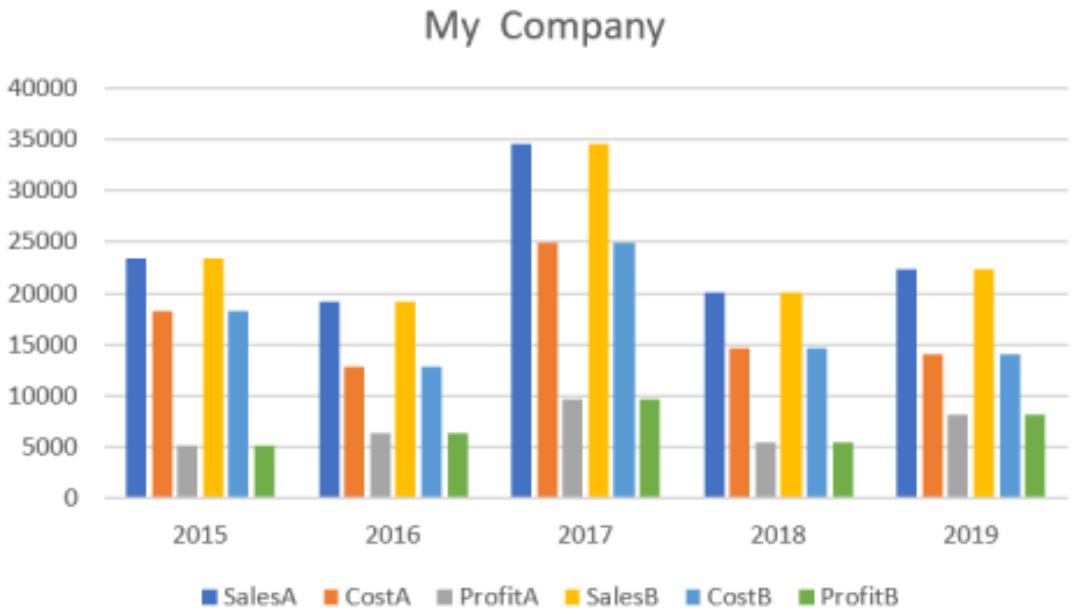
Showing basic comparisons & relationships b/w variables

The basic parts of an Excel diagram

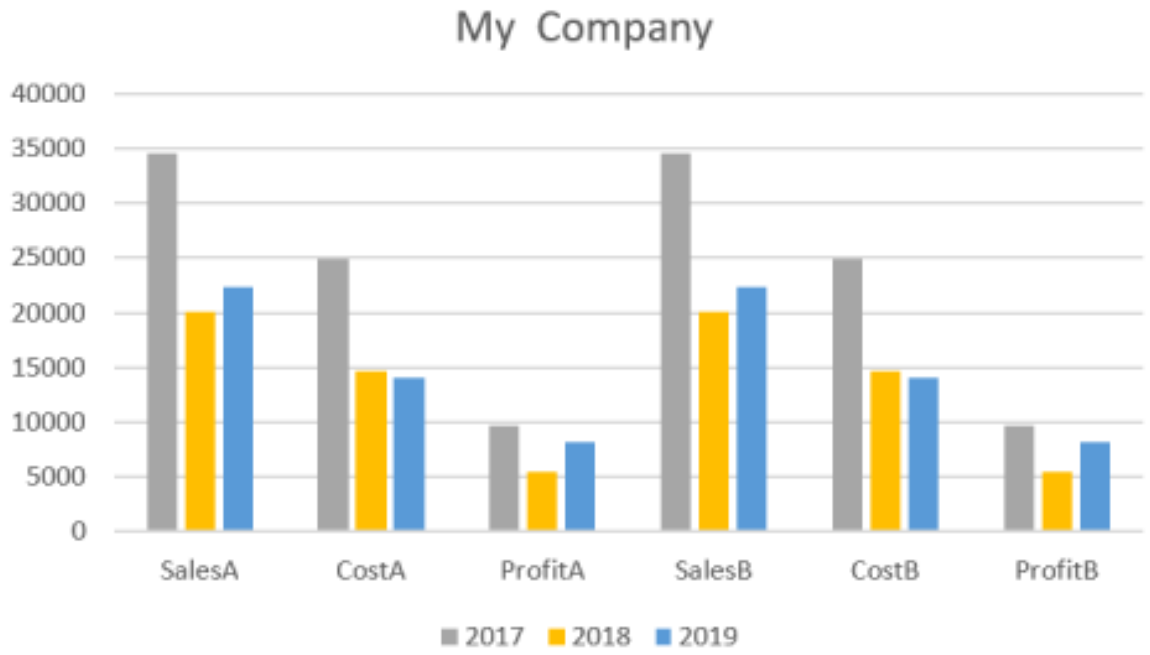


Showing basic comparisons & relationships b/w variables

Column charts

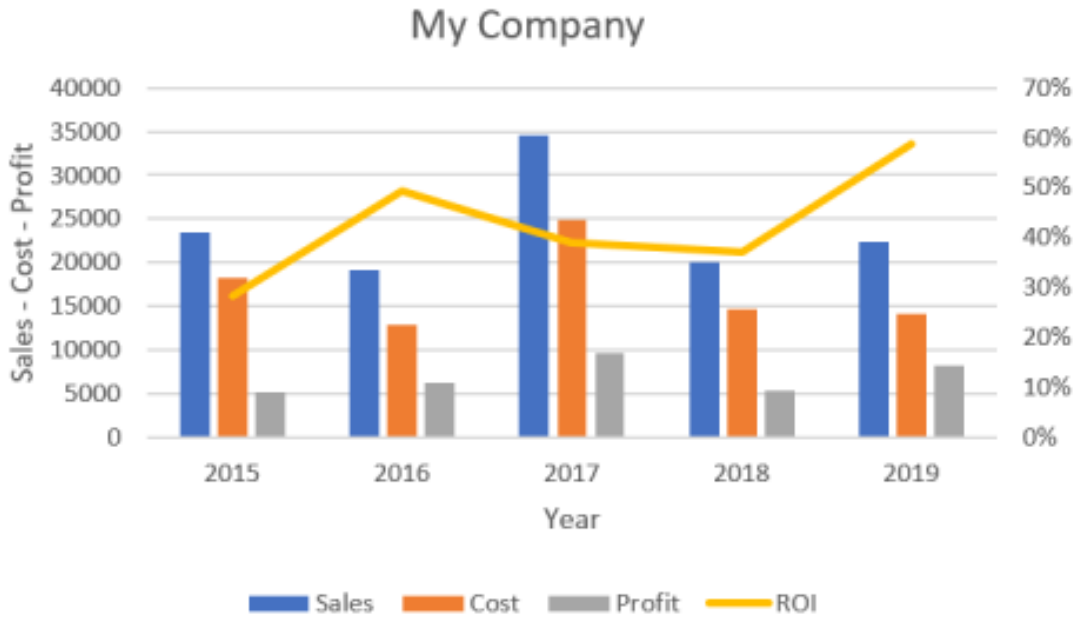


Showing basic comparisons & relationships b/w variables



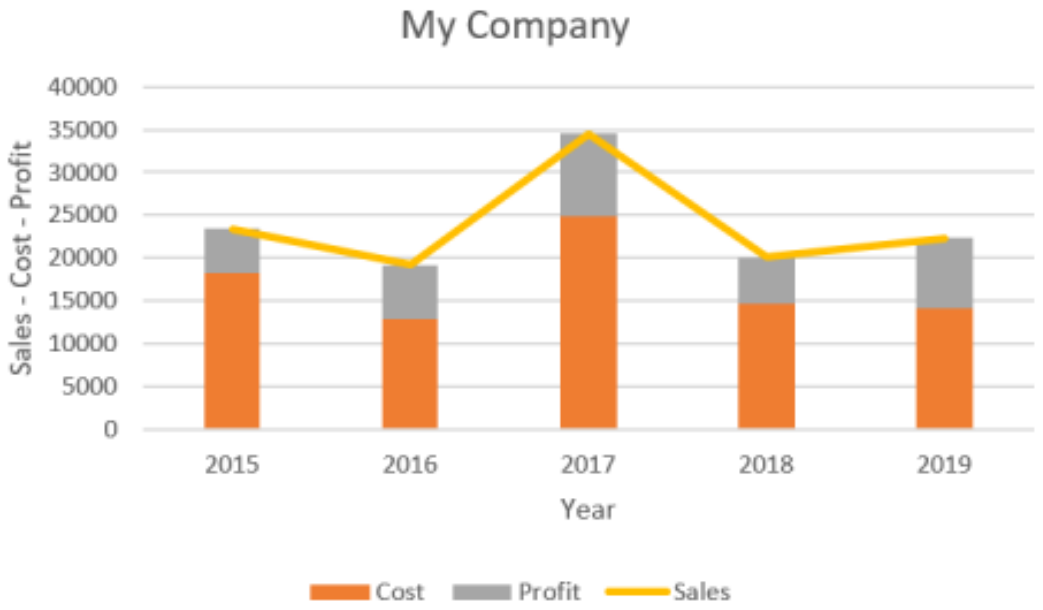
Showing basic comparisons & relationships b/w variables

Combination charts



Showing basic comparisons & relationships b/w variables

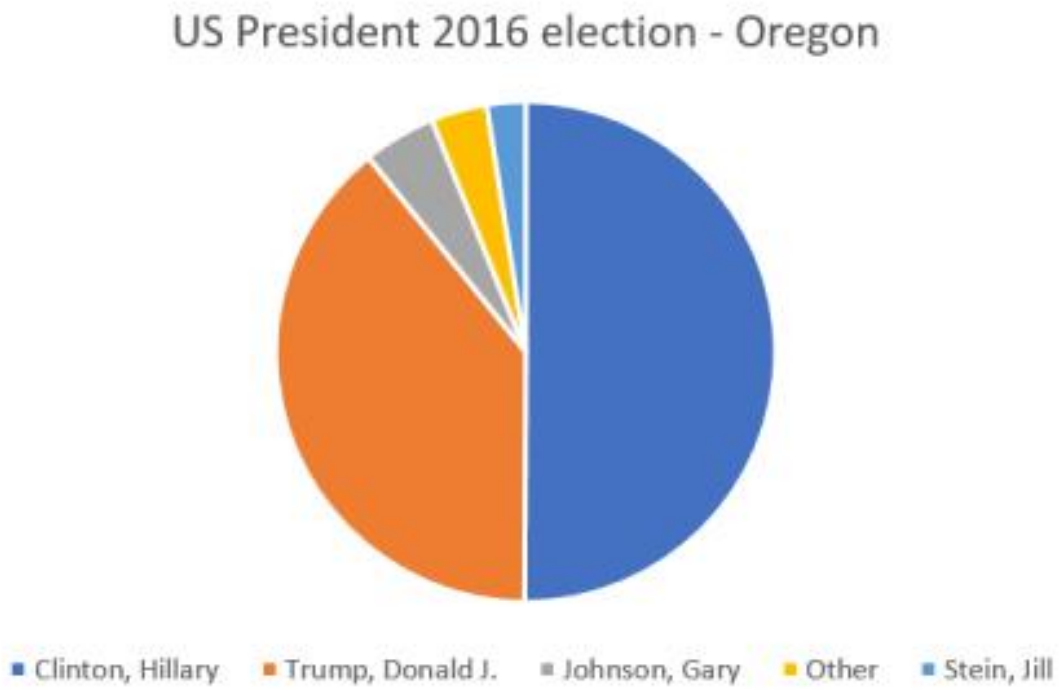
Stacked charts



Pie and bar charts

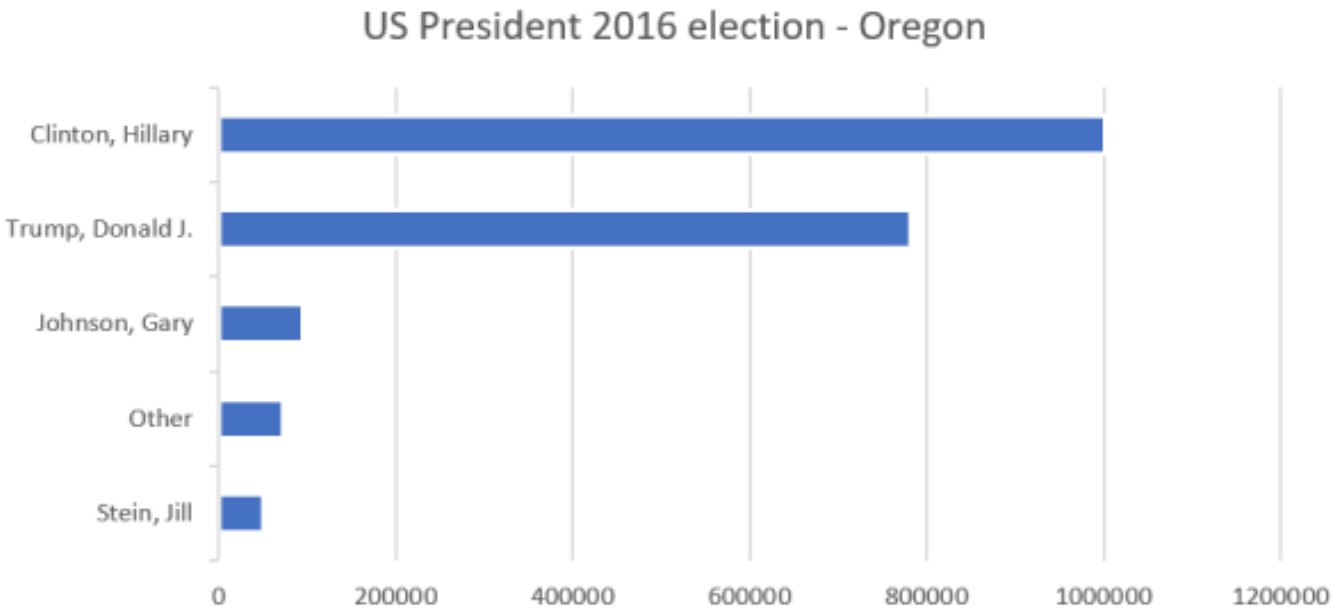
Candidate	Votes
Clinton, Hillary	1,002,106
Trump, Donald J.	782,403
Johnson, Gary	94,231
Other	72,594
Stein, Jill	50,002

Showing basic comparisons & relationships b/w variables

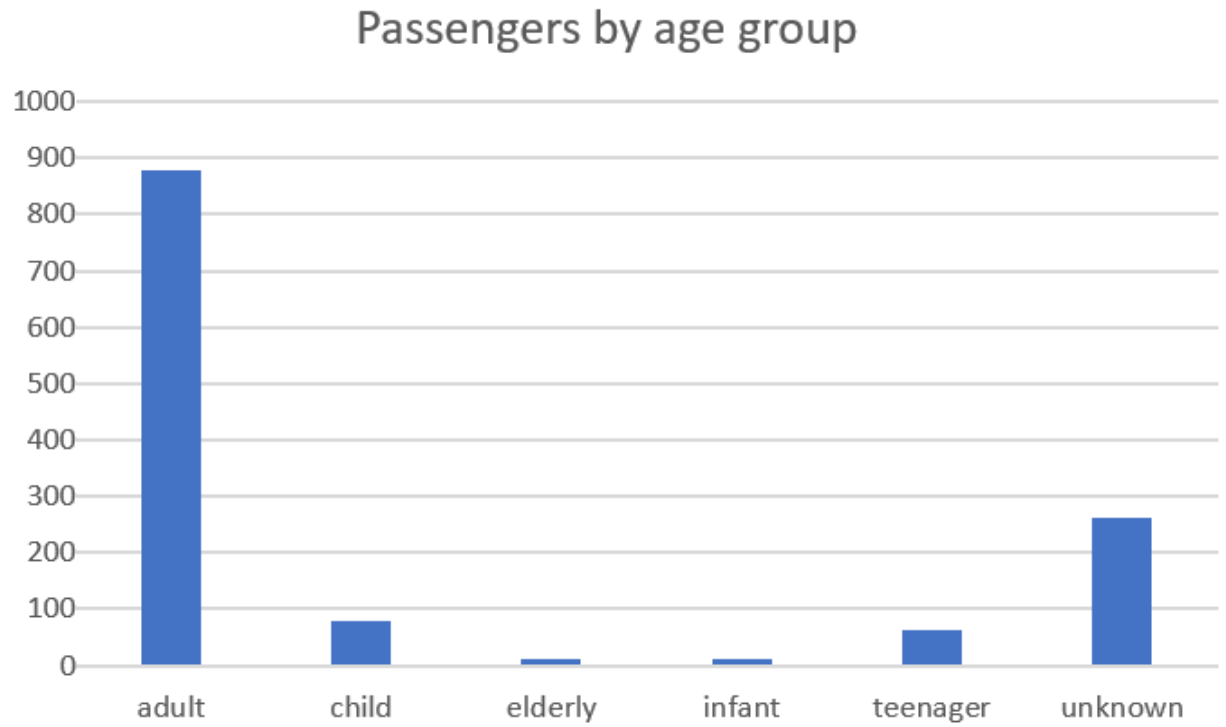


Showing basic comparisons & relationships b/w variables

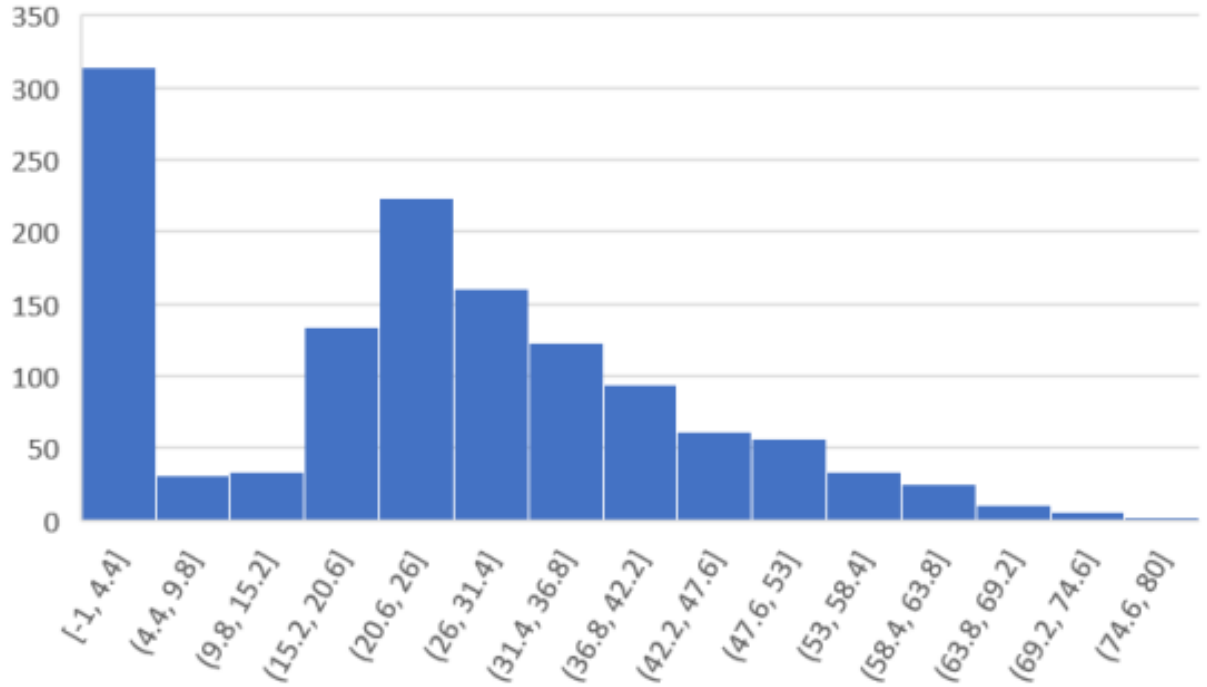
- The bar chart will look similar to the following:



Building data distributions using histograms



Building data distributions using histograms

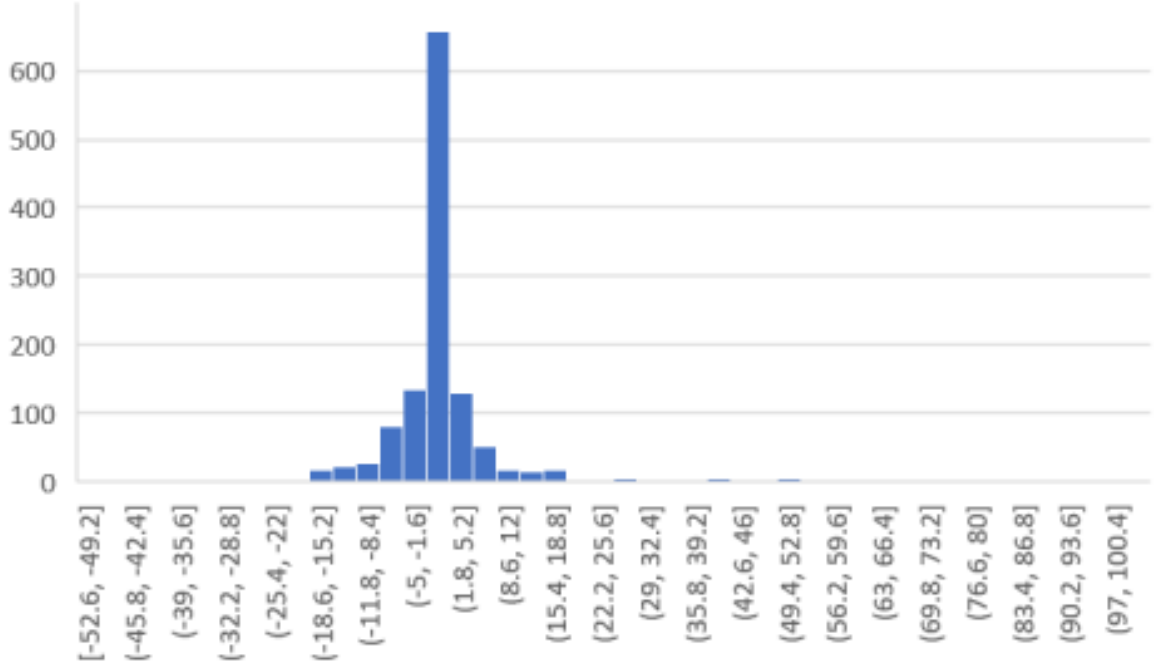


Building data distributions using histograms

The preceding histogram was created following these steps:

- Navigate to Insert | Histogram.
- Double-click the x axis to set the number of bins to 15.

Building data distributions using histograms



Representing geographical distribution of data in maps

- Localizing information in a map is extremely useful to understand data in the spatial dimension, which is often difficult by other means.
- Excel offers different options and we are going to show a couple of them.
- We will start by using data containing geographical coordinates, that is, latitude and longitude.

Representing geographical distribution of data in maps

- The input data table is the following:

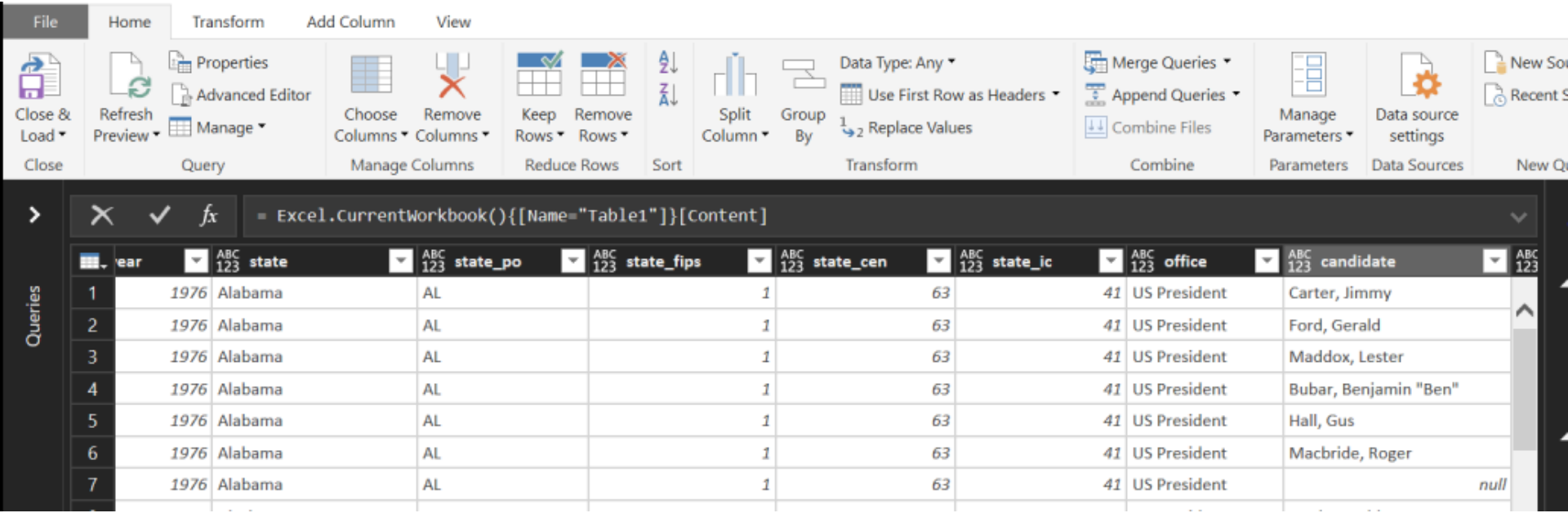
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	year	state	state_po	state_fips	state_cen	state_ic	office	candidate	party	writein	candidatevotes	totalvotes	version
2	1976	Alabama	AL	1	63	41	US President	Carter, Jimmy	democrat	FALSE	659170	1182850	20171015
3	1976	Alabama	AL	1	63	41	US President	Ford, Gerald	republican	FALSE	504070	1182850	20171015
4	1976	Alabama	AL	1	63	41	US President	Maddox, Lester	american independent party	FALSE	9198	1182850	20171015
5	1976	Alabama	AL	1	63	41	US President	Bubar, Benjamin "Ben"	prohibition	FALSE	6669	1182850	20171015
6	1976	Alabama	AL	1	63	41	US President	Hall, Gus	communist party use	FALSE	1954	1182850	20171015
7	1976	Alabama	AL	1	63	41	US President	Macbride, Roger	libertarian	FALSE	1481	1182850	20171015
8	1976	Alabama	AL	1	63	41	US President			TRUE	308	1182850	20171015
9	1976	Alaska	AK	2	94	81	US President	Ford, Gerald	republican	FALSE	71555	123574	20171015
10	1976	Alaska	AK	2	94	81	US President	Carter, Jimmy	democrat	FALSE	44058	123574	20171015
11	1976	Alaska	AK	2	94	81	US President	Macbride, Roger	libertarian	FALSE	6785	123574	20171015
12	1976	Alaska	AK	2	94	81	US President			TRUE	1176	123574	20171015
13	1976	Arizona	AZ	4	86	61	US President	Ford, Gerald	republican	FALSE	418642	742719	20171015
14	1976	Arizona	AZ	4	86	61	US President	Carter, Jimmy	democrat	FALSE	295602	742719	20171015

Representing geographical distribution of data in maps

From this table, we need to extract the winner party in each state, namely the one with more votes. We will use Power Query and its Group By function.

- Navigate to Data | From Table/Range.
- Open the Power Query window. You should see something similar to the following screenshot in next slide.

Representing geographical distribution of data in maps



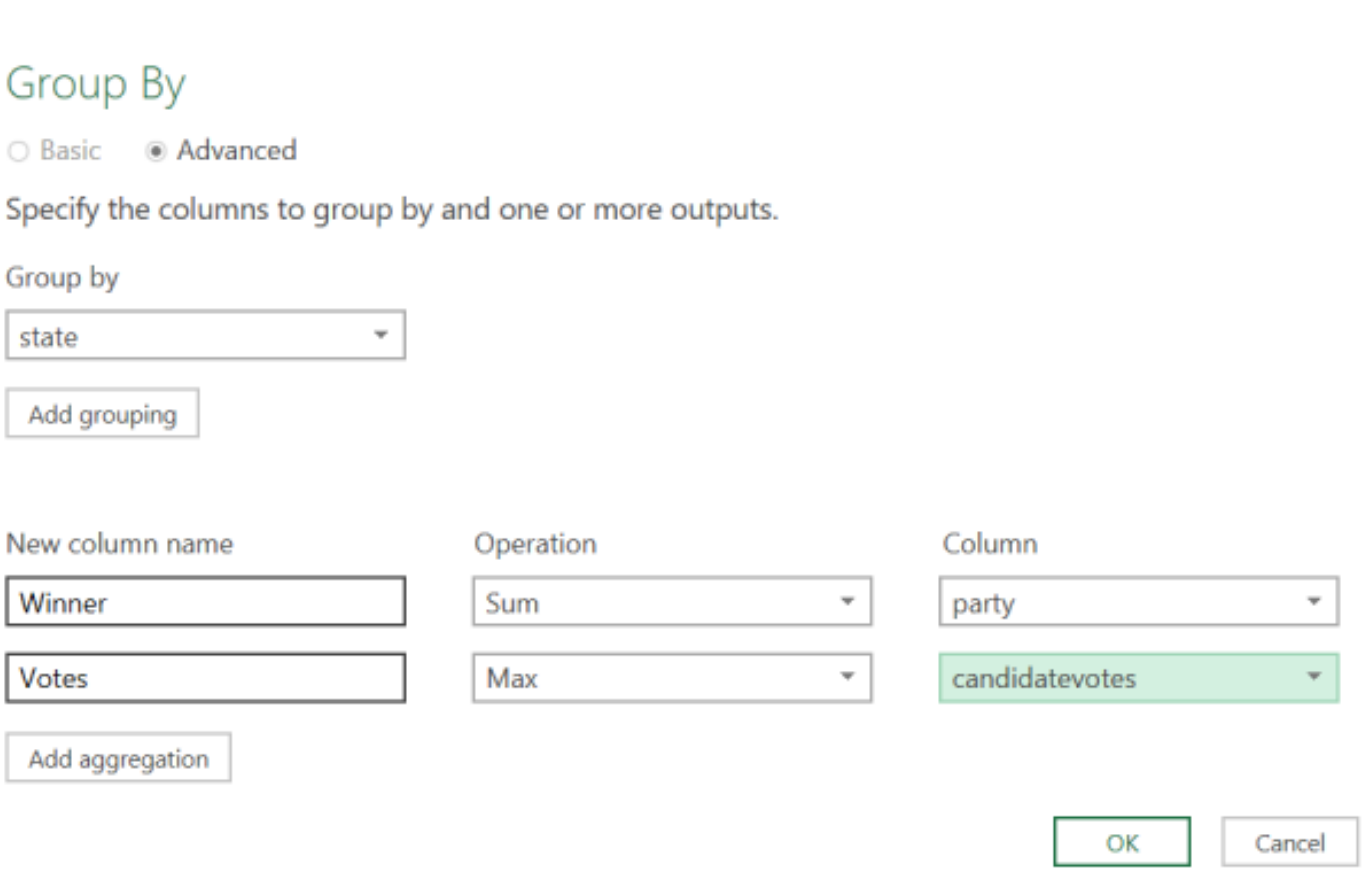
	ear	state	state_po	state_fips	state_cen	state_ic	office	candidate	
1	1976	Alabama	AL	1	63	41	US President	Carter, Jimmy	
2	1976	Alabama	AL	1	63	41	US President	Ford, Gerald	
3	1976	Alabama	AL	1	63	41	US President	Maddox, Lester	
4	1976	Alabama	AL	1	63	41	US President	Bubar, Benjamin "Ben"	
5	1976	Alabama	AL	1	63	41	US President	Hall, Gus	
6	1976	Alabama	AL	1	63	41	US President	Macbride, Roger	
7	1976	Alabama	AL	1	63	41	US President	null	

Representing geographical distribution of data in maps

- Select Group By | Advanced then select state for the Group by option.

- Then, we will add a new column called Winner, where we will Sum the values of party. This will give an error, but will give us the base function to get the winner party name later.

- The second column we will add is named Votes, where we select the maximum value of candidatevotes. This will show the larger number of votes in each state.


Group By
Basic
Advanced
Specify the columns to group by and one or more outputs.
Group by
state
Add grouping
New column name
Operation
Column
Winner
Sum
party
Votes
Max
candidatevotes
Add aggregation
OK
Cancel

Representing geographical distribution of data in maps

Queries

```
= Table.Group(#"Changed Type", {"state"}, {{"Winner", each List.Sum([party]), type text}, {"Votes", each List.Max
```

	state	Winner	Votes
1	Alabama	Error	659170
2	Alaska	Error	71555
3	Arizona	Error	418642
4	Arkansas	Error	498604
5	California	Error	3882244
6	Colorado	Error	584278
7	Connecticut	Error	719261
8	Delaware	Error	122461
9	District of Columbia	Error	137818
10	Florida	Error	1636000
11	Georgia	Error	979409

Representing geographical distribution of data in maps

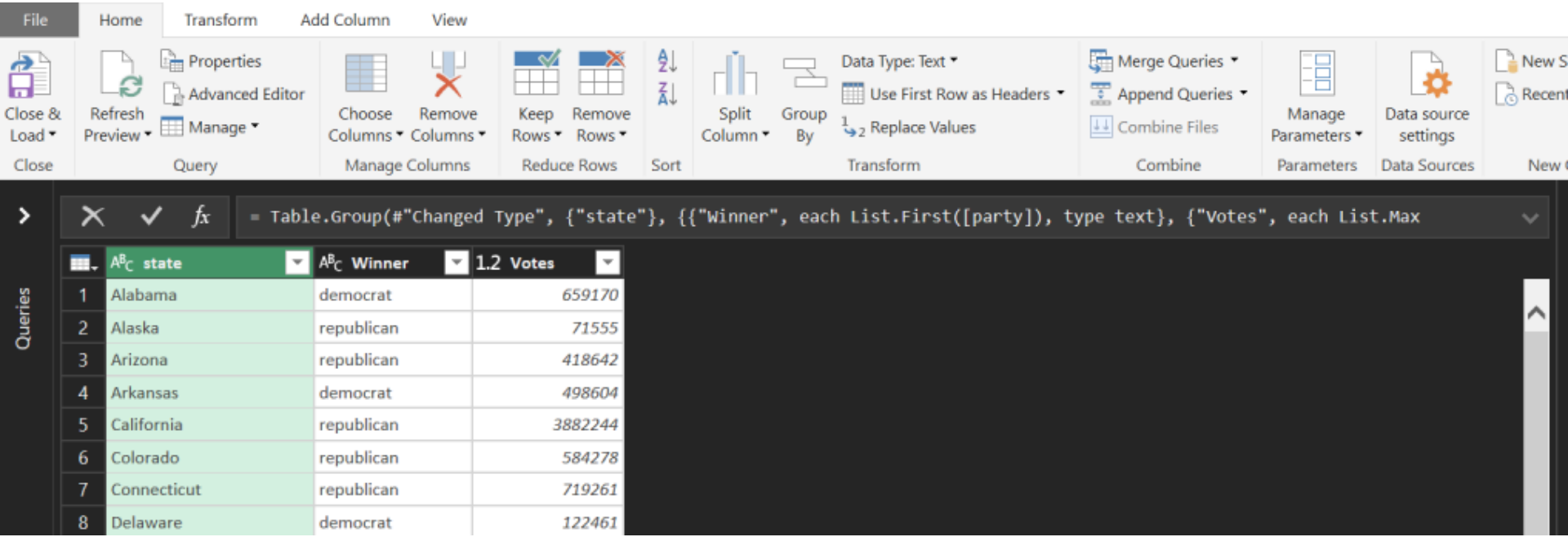
- To fix the error and get the party name in the Winner column, we replace the function:

*= Table.Group(#"Changed Type", {"state"}, {{"Winner", each List.**Sum**([party]), type text}, {"Votes", each List.Max([candidatevotes]), type number}})*

We replace the preceding function with the following:

*= Table.Group(#"Changed Type", {"state"}, {{"Winner", each List.**First**([party]), type text}, {"Votes", each List.Max([candidatevotes]), type number}})*

Representing geographical distribution of data in maps

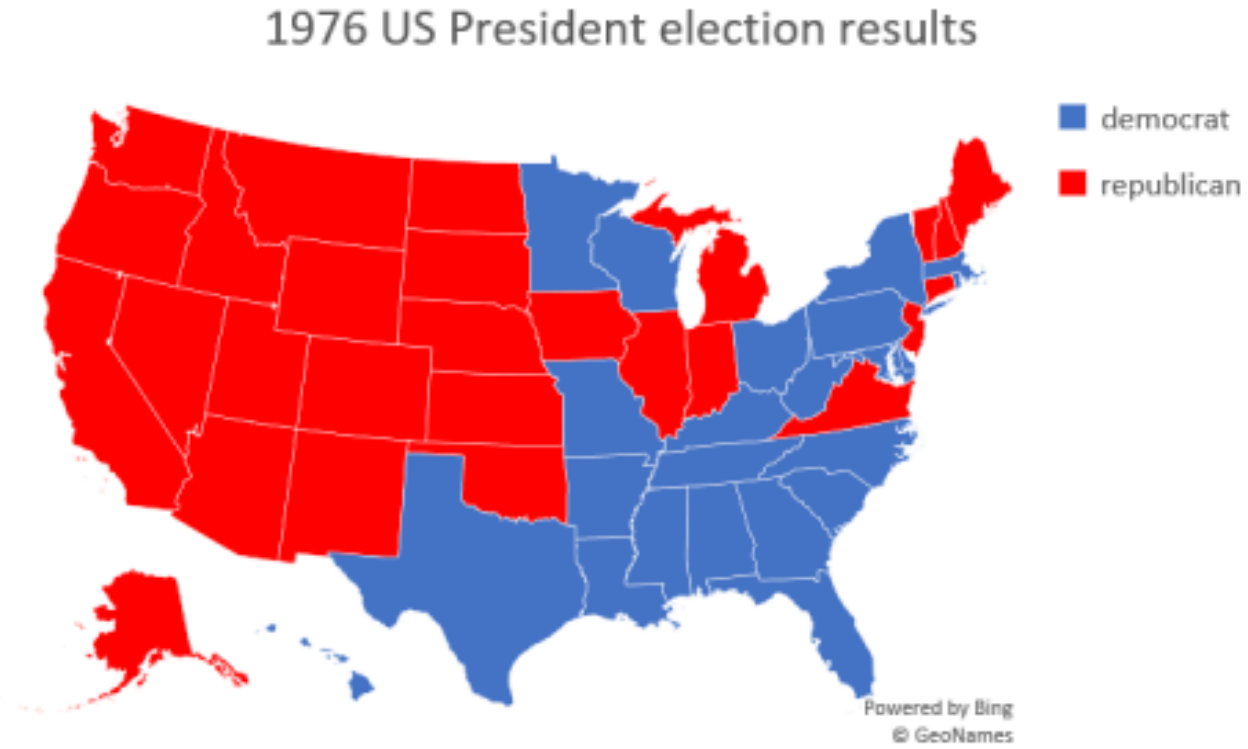


	state	Winner	Votes
1	Alabama	democrat	659170
2	Alaska	republican	71555
3	Arizona	republican	418642
4	Arkansas	democrat	498604
5	California	republican	3882244
6	Colorado	republican	584278
7	Connecticut	republican	719261
8	Delaware	democrat	122461

Representing geographical distribution of data in maps

- Click on Close & Load.
- Use the generated table to create the map. Select any cell in the table and then navigate to Insert | Recommended Charts. The first suggestion will be a map like the one we want.
- Click OK.

Representing geographical distribution of data in maps



Representing geographical distribution of data in maps

- Our second example will use data from the Argentinean Government public database (datos.gob.ar).
- The latitude and longitude of all underground stations in Buenos Aires is listed, together with their names and the lines they belong to.
- The nickname of the Buenos Aires underground is Subte, a short form of the word subterráneo (underground), hence the name of the file.

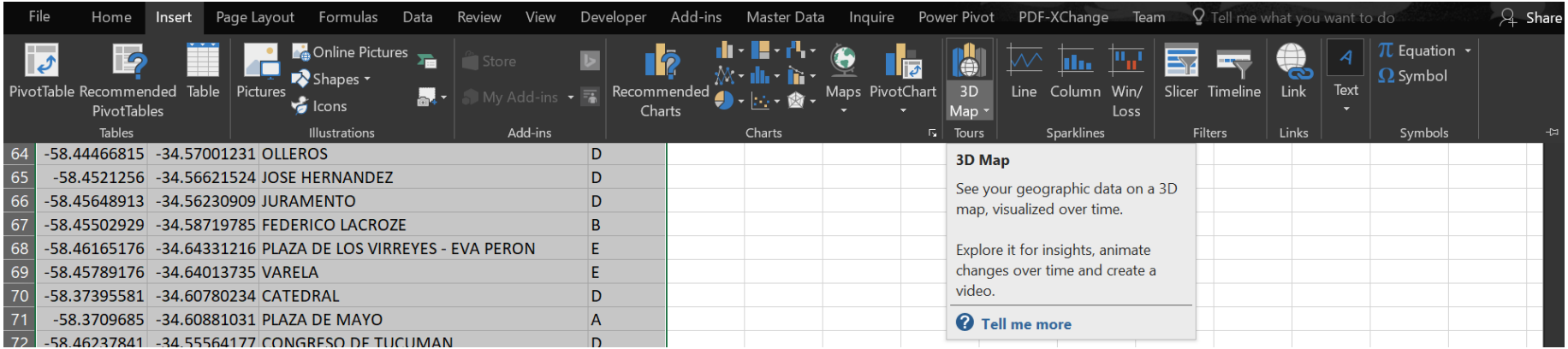
- Load the subte.xlsx file and you will see the following table (partially shown):

	A	B	C	D
1	Long	Lat	Station	Line
2	-58.39892759	-34.63575018	CASEROS	H
3	-58.40096956	-34.62937566	INCLAN - MEZQUITA AL AHMAD	H
4	-58.40232273	-34.62309232	HUMBERTO 1°	H
5	-58.40473172	-34.61524215	VENEZUELA	H
6	-58.40603638	-34.60893524	ONCE - 30 DE DICIEMBRE	H
7	-58.38057434	-34.6042452	9 DE JULIO	D
8	-58.39792376	-34.59975708	FACULTAD DE MEDICINA	D
9	-58.38514236	-34.60158717	TRIBUNALES - TEATRO COLÓN	D
10	-58.40716132	-34.59162784	AGÜERO	D
11	-58.41595542	-34.58515594	R.SCALABRINI ORTIZ	D
12	-58.42119601	-34.58141119	PLAZA ITALIA	D
13	-58.42571144	-34.57842202	PALERMO	D
14	-58.37401822	-34.59119381	RETIRO	C
15	-58.37815578	-34.60176992	LAVALLE	C
16	-58.37952998	-34.60484374	DIAGONAL NORTE	C
17	-58.38061072	-34.60898331	AV. DE MAYO	C
18	-58.38044447	-34.61261728	MORENO	C
19	-58.38017361	-34.6181256	INDEPENDENCIA	C
20	-58.38143443	-34.62761945	CONSTITUCION	C
21	-58.37507152	-34.60329729	FLORIDA	B
22	-58.38071485	-34.60363711	C. PELLEGRINI	B
23	-58.38729613	-34.60409355	URUGUAY	B
24	-58.39231424	-34.60441954	CALLAO - MAESTRO ALFREDO BRAVO	B
25	-58.39947426	-34.60464297	PASTEUR - AMIA	B
26	-58.40539944	-34.60458106	PUEYRREDON	B
27	-58.4117626	-34.60407952	CARLOS GARDEL	B

Underground Stations

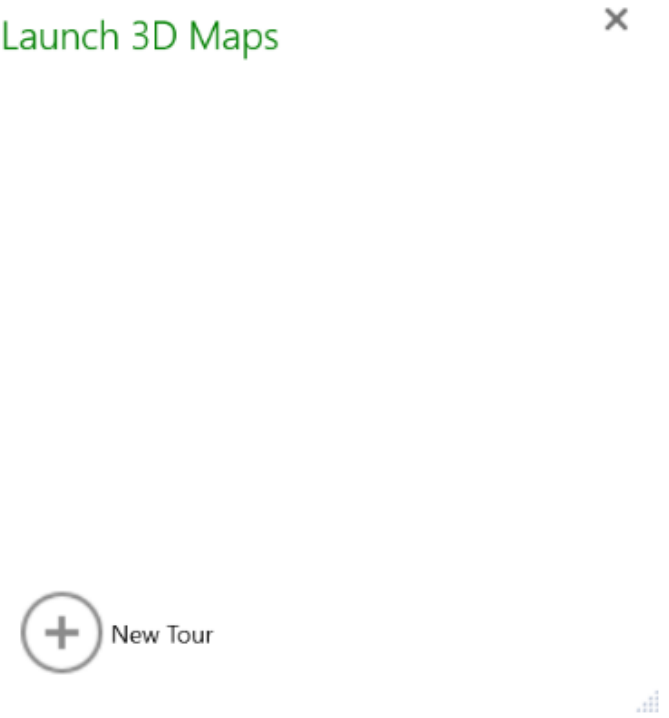
Representing geographical distribution of data in maps

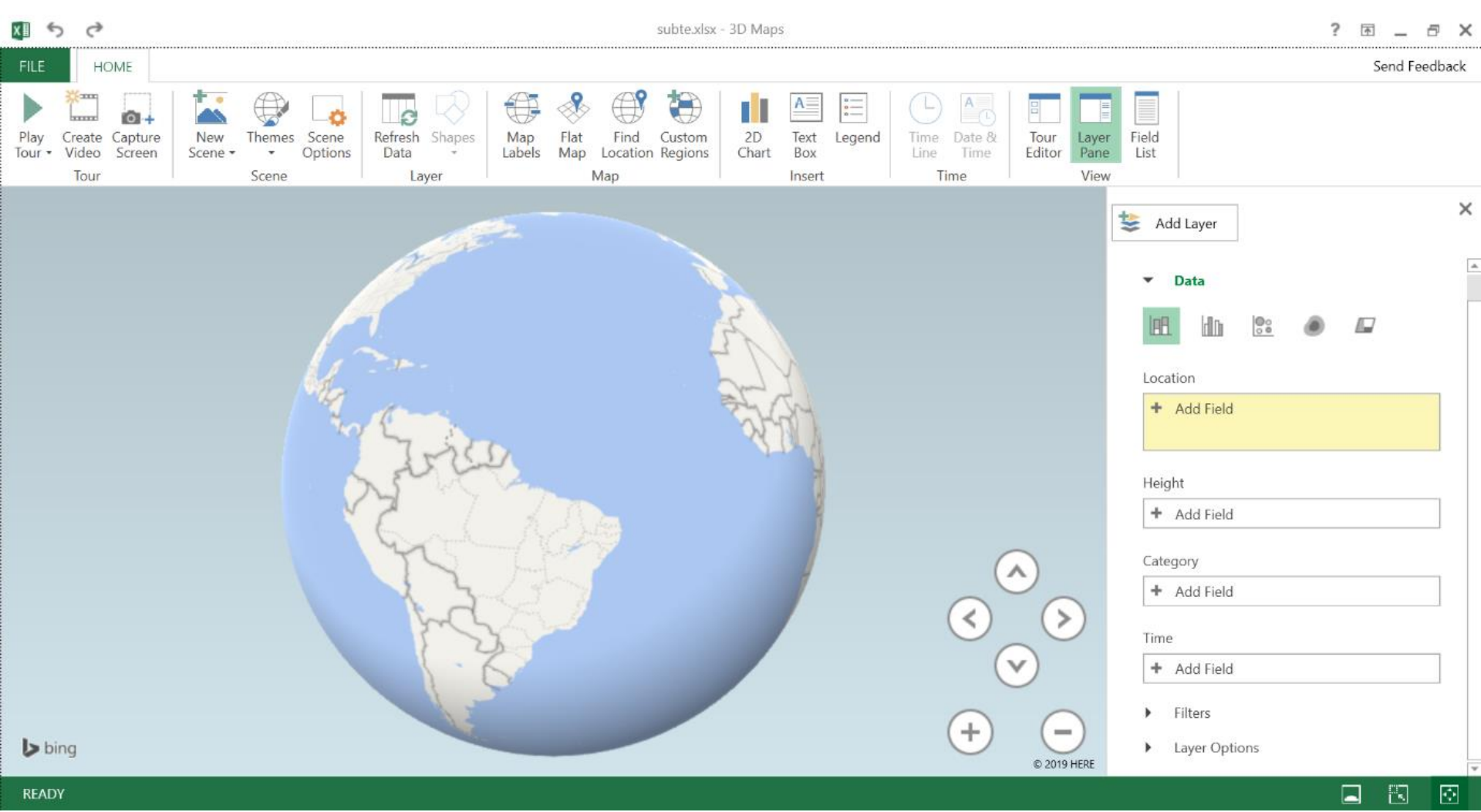
- Select the full data range.
- Navigate to Insert | 3D Map. You should see something similar to the following screenshot:



Representing geographical distribution of data in maps

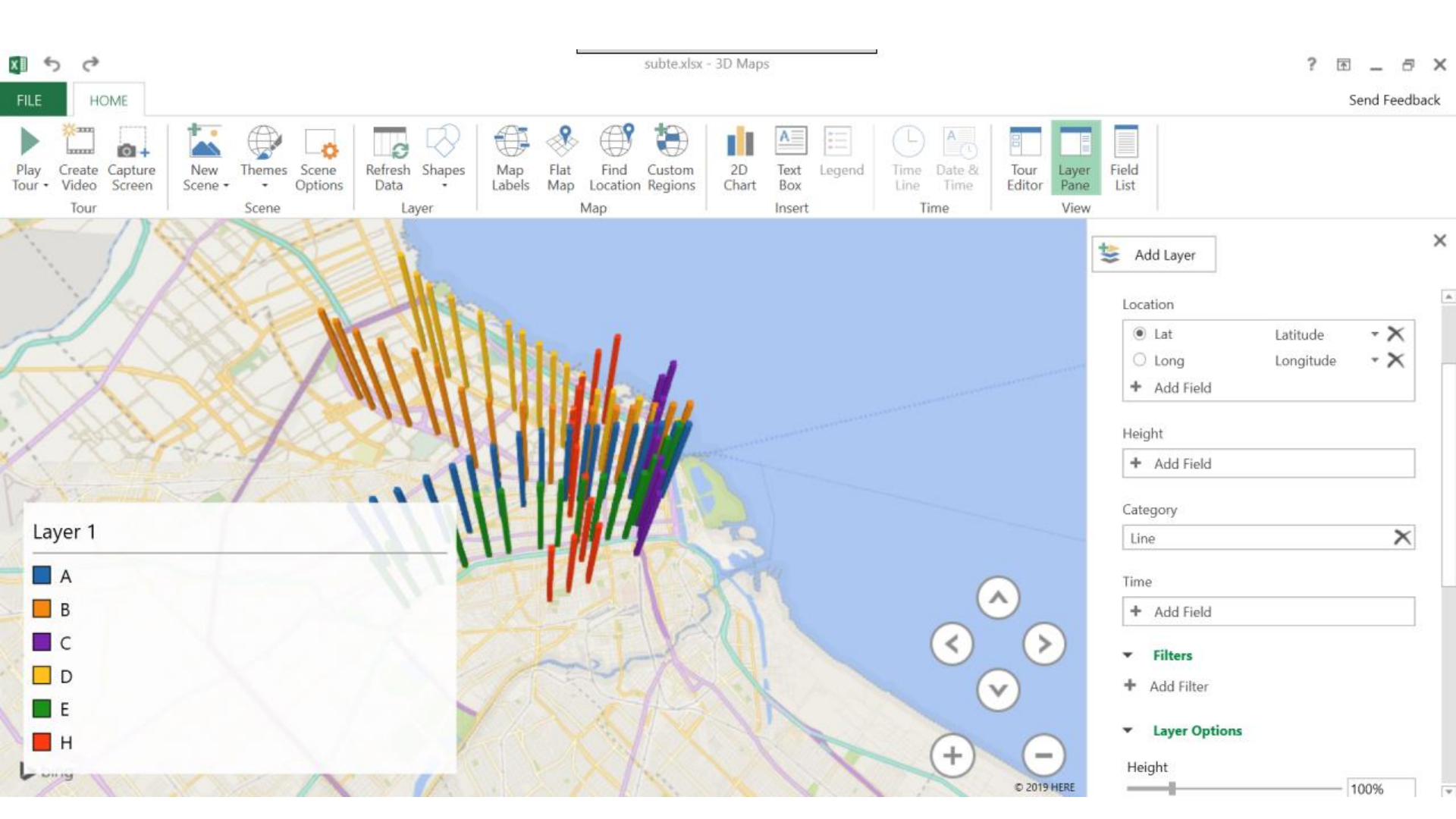
- Click on New tour as shown in the following screenshot:




subte.xlsx - 3D Maps
FILE
HOME
Send Feedback
Play Tour
Create Video
Capture Screen
Tour
New Scene
Themes
Scene Options
Scene
Refresh Data
Shapes
Layer
Map Labels
Flat Map
Find Location
Custom Regions
Map
2D Chart
Text Box
Legend
Insert
Time Line
Date & Time
Time
Tour Editor
Layer Pane
Field List
View
Add Layer
Data
Location
Add Field
Height
Add Field
Category
Add Field
Time
Add Field
Filters
Layer Options
bing
© 2019 HERE
READY

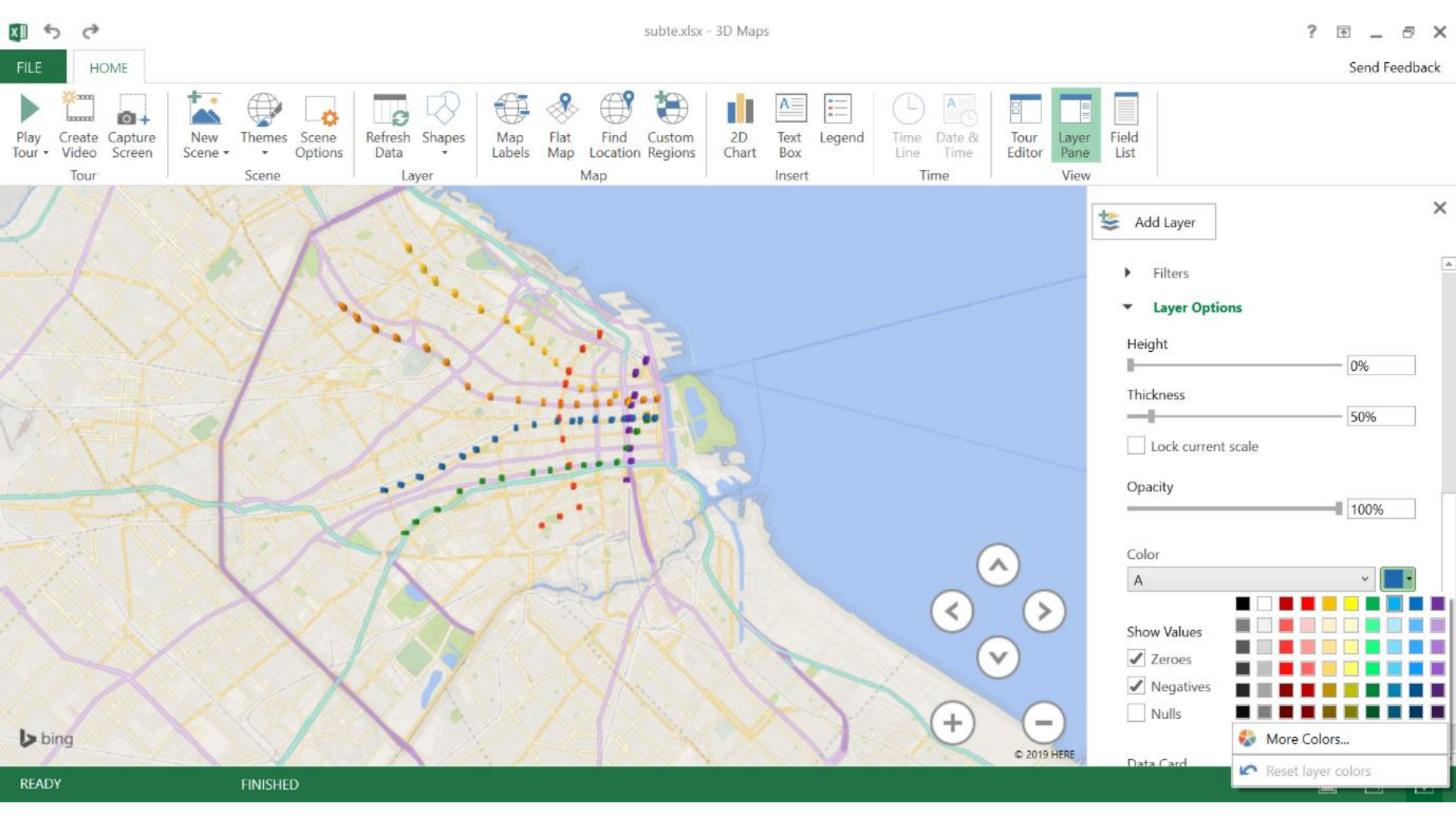
Representing geographical distribution of data in maps

- Add two fields to the Location window: Long and Lat. They should be automatically assigned to the corresponding variables. If not, select them from the menu to the right of the names (see the following screenshot).
- The map should now be centered in the city of Buenos Aires, but the zoom might still be too far out.
- Zoom in by scrolling with your mouse or using the + button.


subte.xlsx - 3D Maps
FILE
HOME
Send Feedback
Play Tour
Create Video
Capture Screen
Tour
New Scene
Themes
Scene Options
Scene
Refresh Data
Shapes
Layer
Map Labels
Flat Map
Find Location
Custom Regions
Map
2D Chart
Text Box
Legend
Insert
Time Line
Date & Time
Time
Tour Editor
Layer Pane
Field List
View
Layer 1
A
B
C
D
E
H
© 2019 HERE
Add Layer
Location
Lat
Latitude
Long
Longitude
Add Field
Height
Add Field
Category
Line
Time
Add Field
Filters
Add Filter
Layer Options
Height
100%

Representing geographical distribution of data in maps

- Click on Layer Options.
- Set Height to 0% and Thickness to between 50% and 70%.
- You can also change the color for each line (category). As in many other cities, the Subte lines are identified by colors.
- A is light blue, B is red, C is blue, D is green, E is purple, and H is yellow (what happened to F and G? They, and I, are not yet built).
- The final result is shown in the following screenshot in next slide


subte.xlsx - 3D Maps
FILE
HOME
Send Feedback
Play Tour
Create Video
Capture Screen
Tour
New Scene
Themes
Scene Options
Scene
Refresh Data
Shapes
Layer
Map Labels
Flat Map
Find Location
Custom Regions
Map
2D Chart
Text Box
Legend
Insert
Time Line
Date & Time
Time
Tour Editor
Layer Pane
Field List
View
Add Layer
Filters
Layer Options
Height
0%
Thickness
50%
Lock current scale
Opacity
100%
Color
A
Show Values
Zeroes
Negatives
Nulls
More Colors...
Reset layer colors
Data Card
bing
© 2019 HERE
READY
FINISHED

Customize Data Card

CHOOSE DATA FIELDS FOR CUSTOM TOOLTIP

Lat

Long

Line

+ Add Field

TEMPLATE 1

Lat
Value

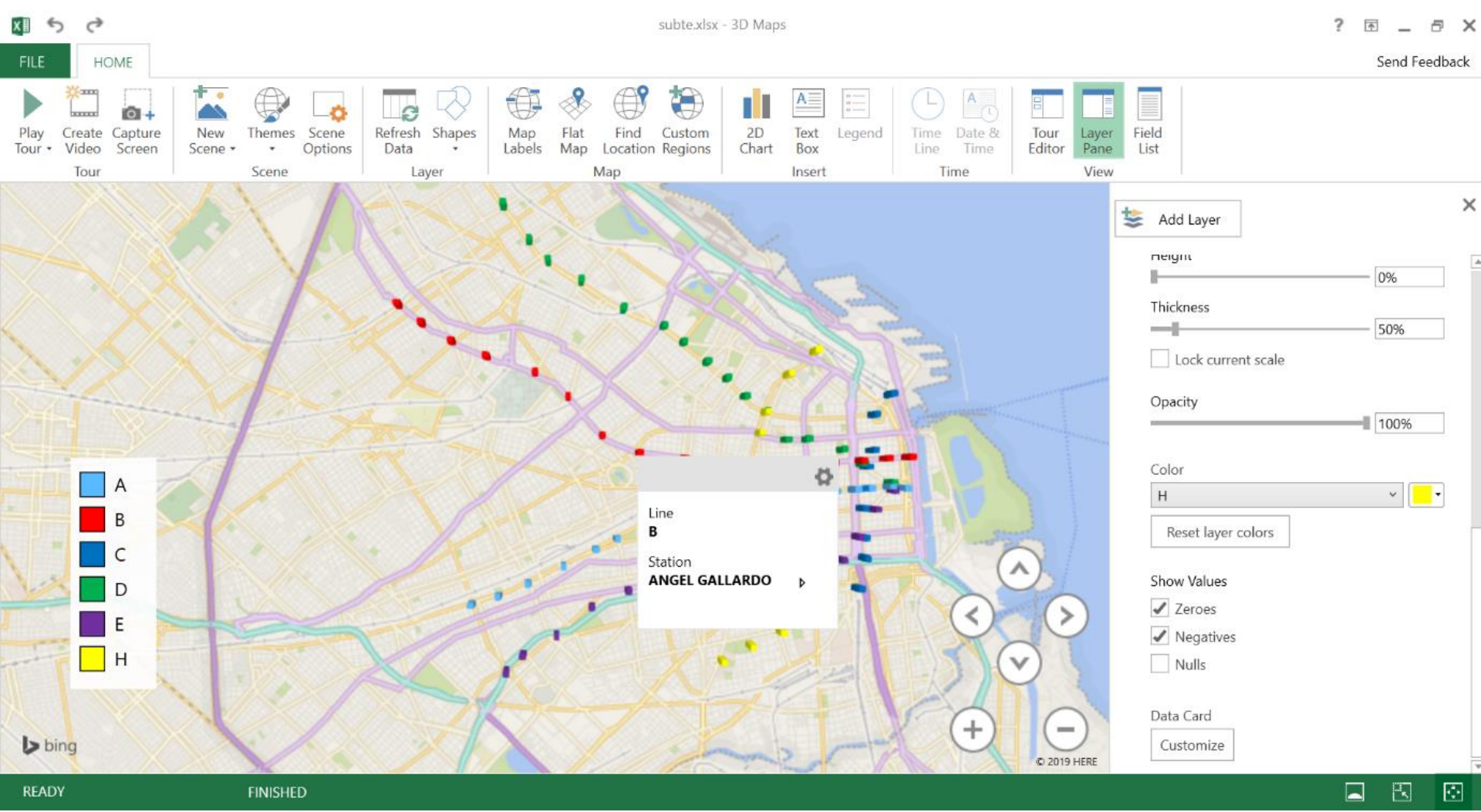
Long
Value

Line
Value

Reset to Defaults

OK

Cancel

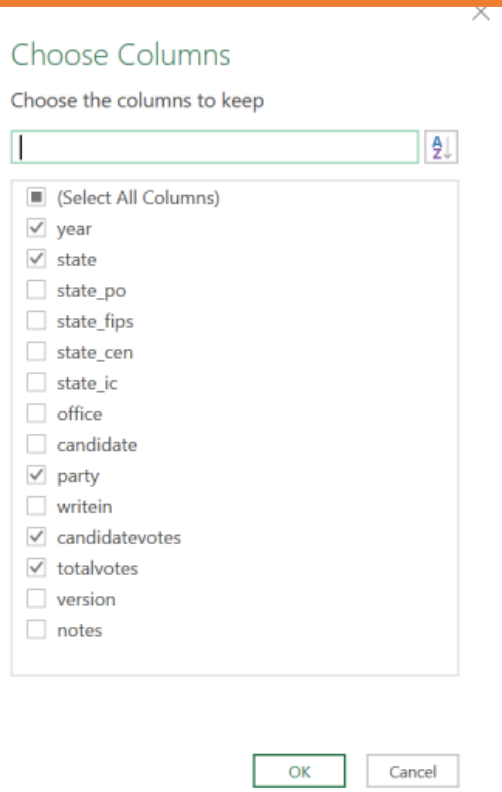

subte.xlsx - 3D Maps
FILE
HOME
Send Feedback
Play Tour
Create Video
Capture Screen
Tour
New Scene
Themes
Scene Options
Scene
Refresh Data
Shapes
Layer
Map Labels
Flat Map
Find Location
Custom Regions
Map
2D Chart
Text Box
Legend
Insert
Time Line
Date & Time
Time
Tour Editor
Layer Pane
Field List
View
A
B
C
D
E
H
Line
B
Station
ANGEL GALLARDO
bing
© 2019 HERE
Add Layer
Height
0%
Thickness
50%
Lock current scale
Opacity
100%
Color
H
Reset layer colors
Show Values
Zeroes
Negatives
Nulls
Data Card
Customize
READY
FINISHED

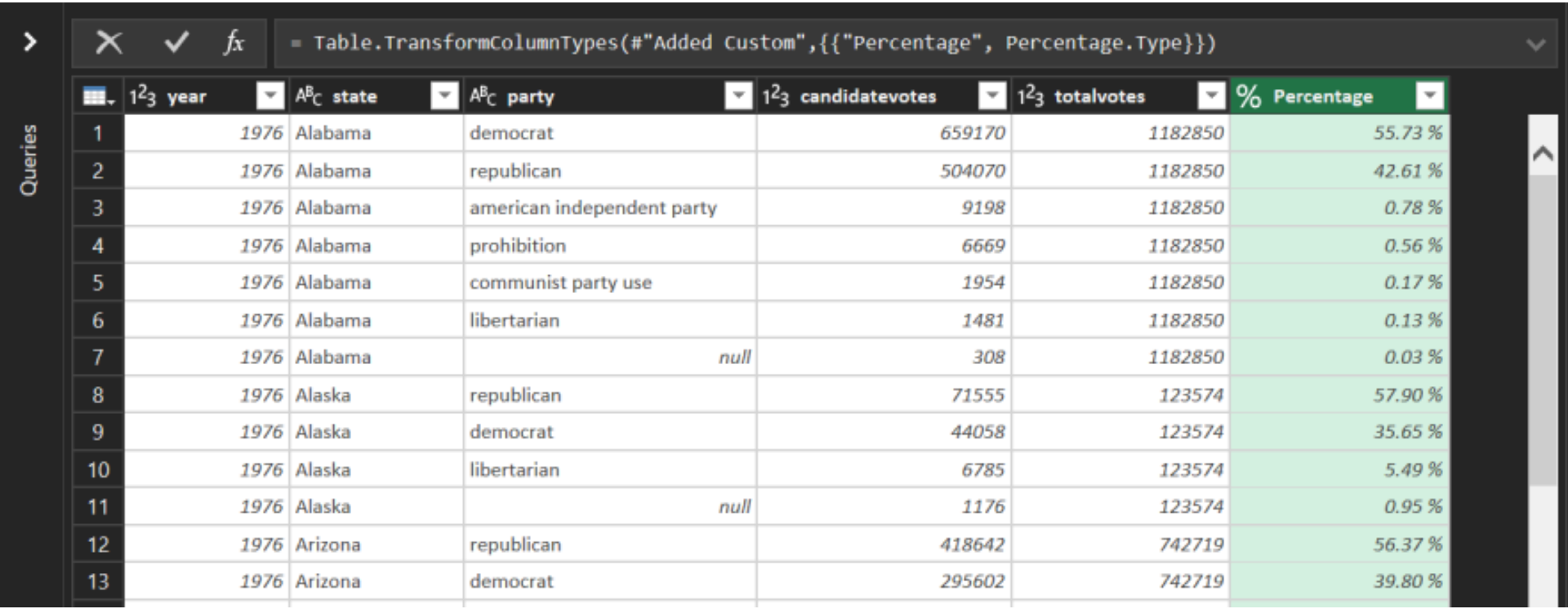
Showing data that changes over time

- Load the 1976_2016USpresident.xlsx file in Excel and you will see the same information in the table that we used in the previous section, except that we can now see the results corresponding to every election in every state from 1976 to 2016.
- Select one state at random (California, in our case) and try to compare how the number of votes per party changed with time.
- Navigate to Data | From Table/Range.
- In the Power Query window, click on Choose Columns.

Showing data that changes over time

- Select only the columns we are interested in: year, state, party, candidatevotes, and totalvotes, as shown in the following screenshot:





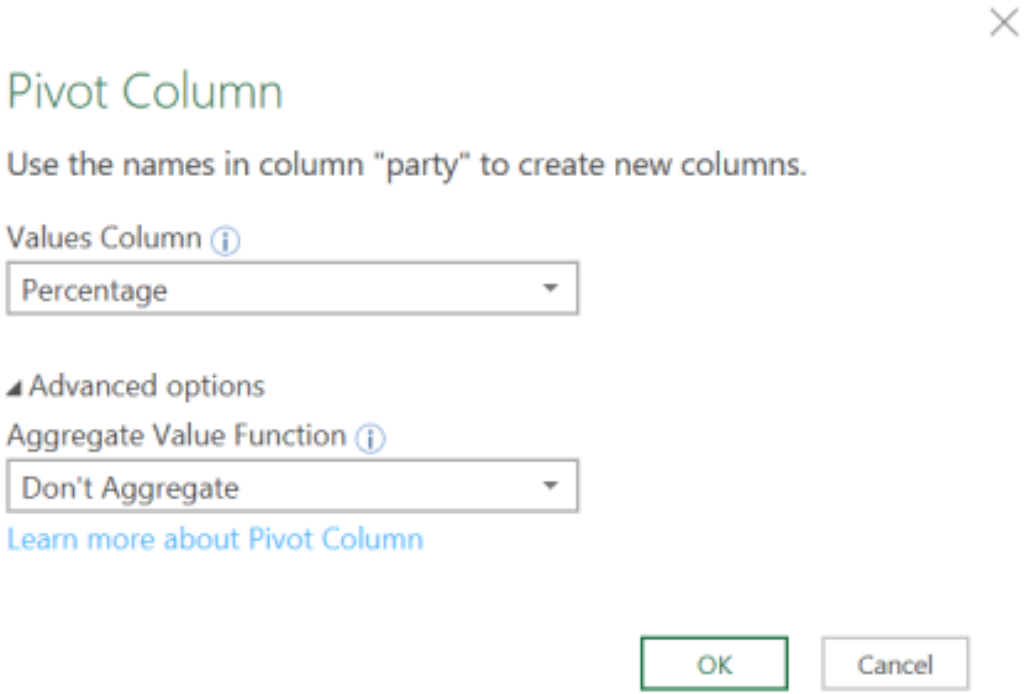
	year	state	party	candidatevotes	totalvotes	Percentage
1	1976	Alabama	democrat	659170	1182850	55.73 %
2	1976	Alabama	republican	504070	1182850	42.61 %
3	1976	Alabama	american independent party	9198	1182850	0.78 %
4	1976	Alabama	prohibition	6669	1182850	0.56 %
5	1976	Alabama	communist party use	1954	1182850	0.17 %
6	1976	Alabama	libertarian	1481	1182850	0.13 %
7	1976	Alabama	null	308	1182850	0.03 %
8	1976	Alaska	republican	71555	123574	57.90 %
9	1976	Alaska	democrat	44058	123574	35.65 %
10	1976	Alaska	libertarian	6785	123574	5.49 %
11	1976	Alaska	null	1176	123574	0.95 %
12	1976	Arizona	republican	418642	742719	56.37 %
13	1976	Arizona	democrat	295602	742719	39.80 %

Showing data that changes over time

We need to format the table in such a way that we can compare both time series in a meaningful way. To do this, perform the following steps:

- Select the party column.
- Navigate to Transform.
- Click on Pivot Column. The pop-up window should look like the following screenshot before clicking OK

Showing data that changes over time



Showing data that changes over time

Queries

```
= Table.Pivot(#"Filtered Rows", List.Distinct(#"Filtered Rows"[party]), "party", "Percentage")
```

	year	state	candidatevotes	totalvotes	republican	democrat
1	1976	California	3742284	7803770	null	47.95 %
2	1976	California	3882244	7803770	49.75 %	null
3	1980	California	3082943	8582938	null	35.92 %
4	1980	California	4522994	8582938	52.70 %	null
5	1984	California	3922519	9505041	null	41.27 %
6	1984	California	5467009	9505041	57.52 %	null
7	1988	California	4702233	9887065	null	47.56 %
8	1988	California	5054917	9887065	51.13 %	null
9	1992	California	3630574	11131721	32.61 %	null
10	1992	California	5121325	11131721	null	46.01 %
11	1996	California	3828381	10019469	38.21 %	null
12	1996	California	5119835	10019469	null	51.10 %

Showing data that changes over time

- Select both % columns.
- Navigate to Transform | Replace Values.
- Change all nulls to zeroes.
- Use Home | Group By to group by state.
- Select Sum as aggregation.

Showing data that changes over time

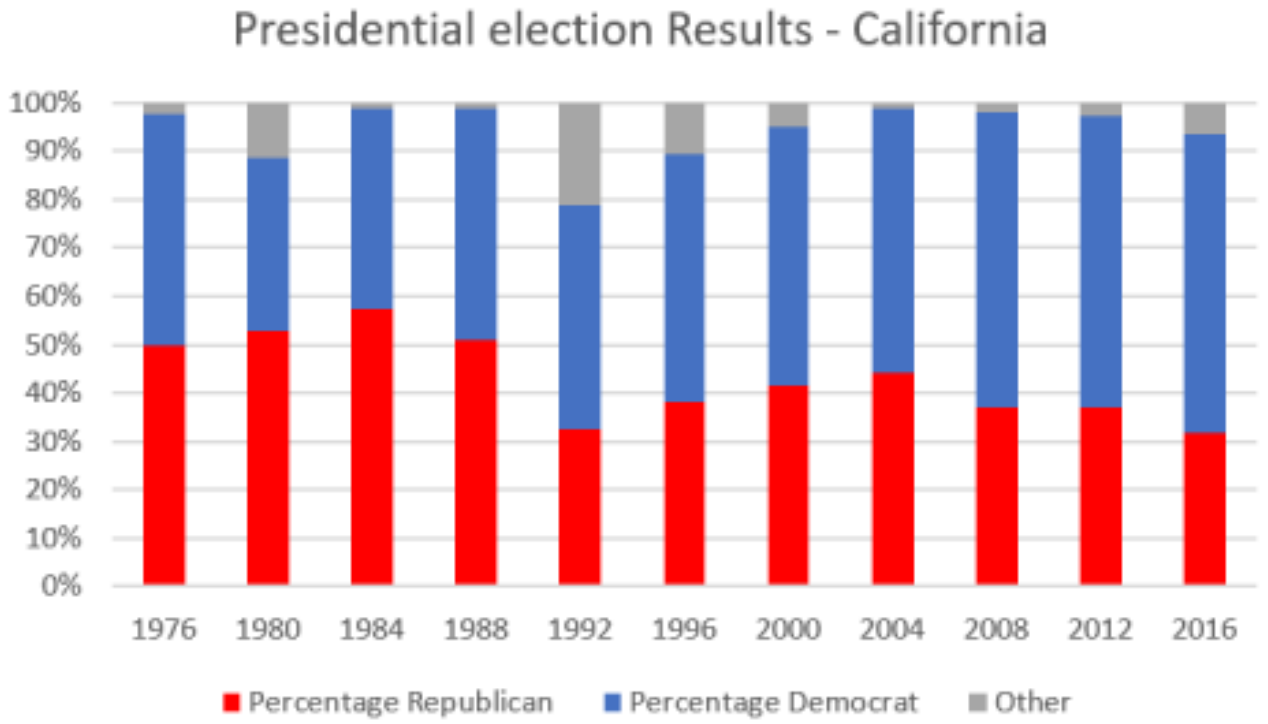
- The last step would be to create another column to account for the difference between the percentage of Democrat and Republican votes and the total:

100%- % republican -% democrat

Showing data that changes over time

	A	B	C	D
1	year	Percentage Republican	Percentage Democrat	Other
2	1976	49.75%	47.95%	2.30%
3	1980	52.70%	35.92%	11.38%
4	1984	57.52%	41.27%	1.22%
5	1988	51.13%	47.56%	1.31%
6	1992	32.61%	46.01%	21.38%
7	1996	38.21%	51.10%	10.69%
8	2000	41.65%	53.45%	4.90%
9	2004	44.36%	54.31%	1.34%
10	2008	36.95%	61.01%	2.03%
11	2012	37.12%	60.24%	2.64%
12	2016	31.62%	61.73%	6.66%

Showing data that changes over time



Summary

- We have discussed different types of diagrams in Excel, which can be used to compare variables and show data in meaningful ways, helping us to extract value from our results.

- We can now go back to the pure machine learning models and take a leap forward to the advanced world of neural networks.

11: The Future of Machine Learning

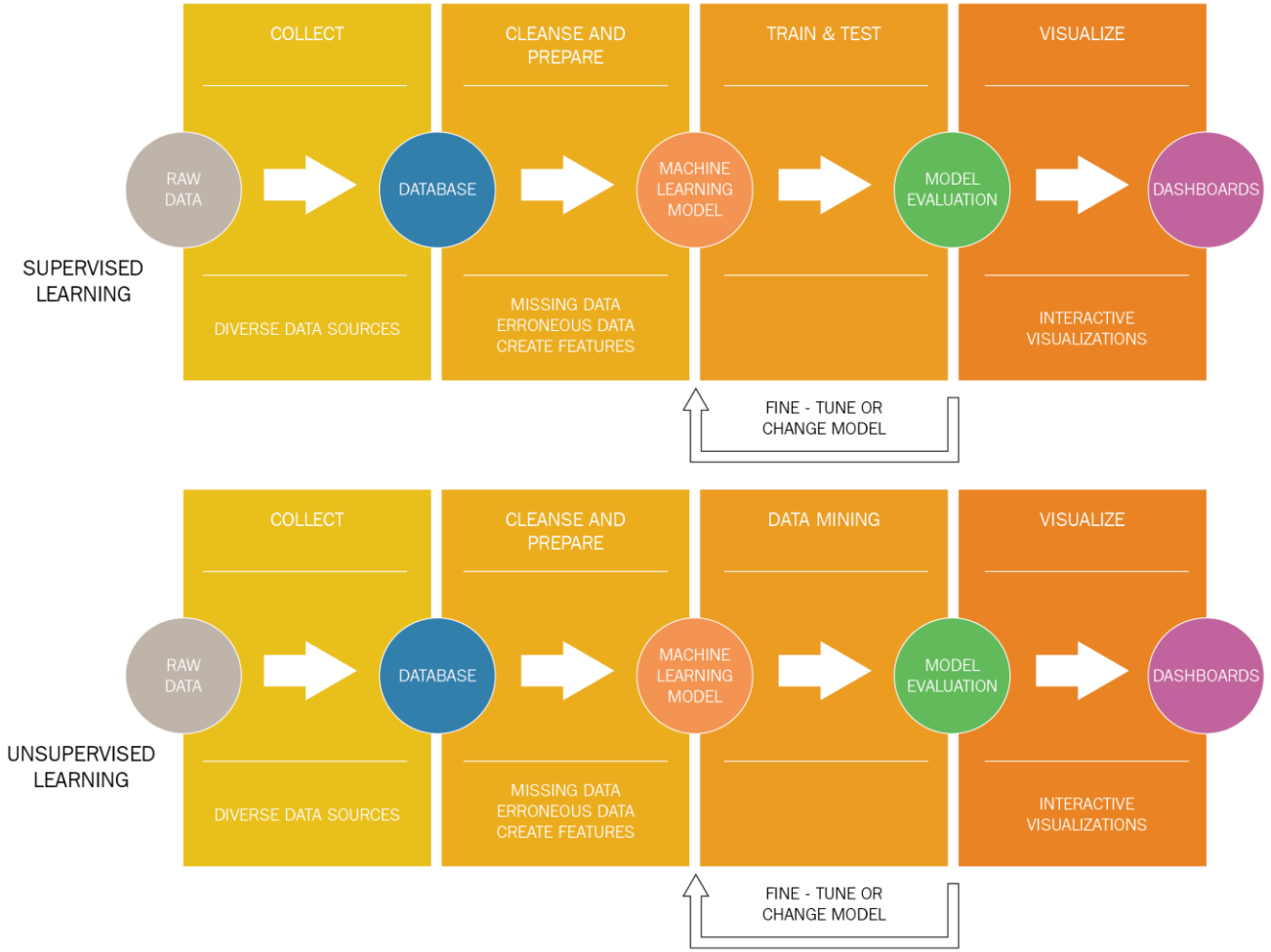
The Future of Automation

The following topics will be covered in this lesson:

- Automatic data analysis flows
- Re-training of the models
- Automatic the analysis
- What can we expect in the future?

Automatic data analysis flows

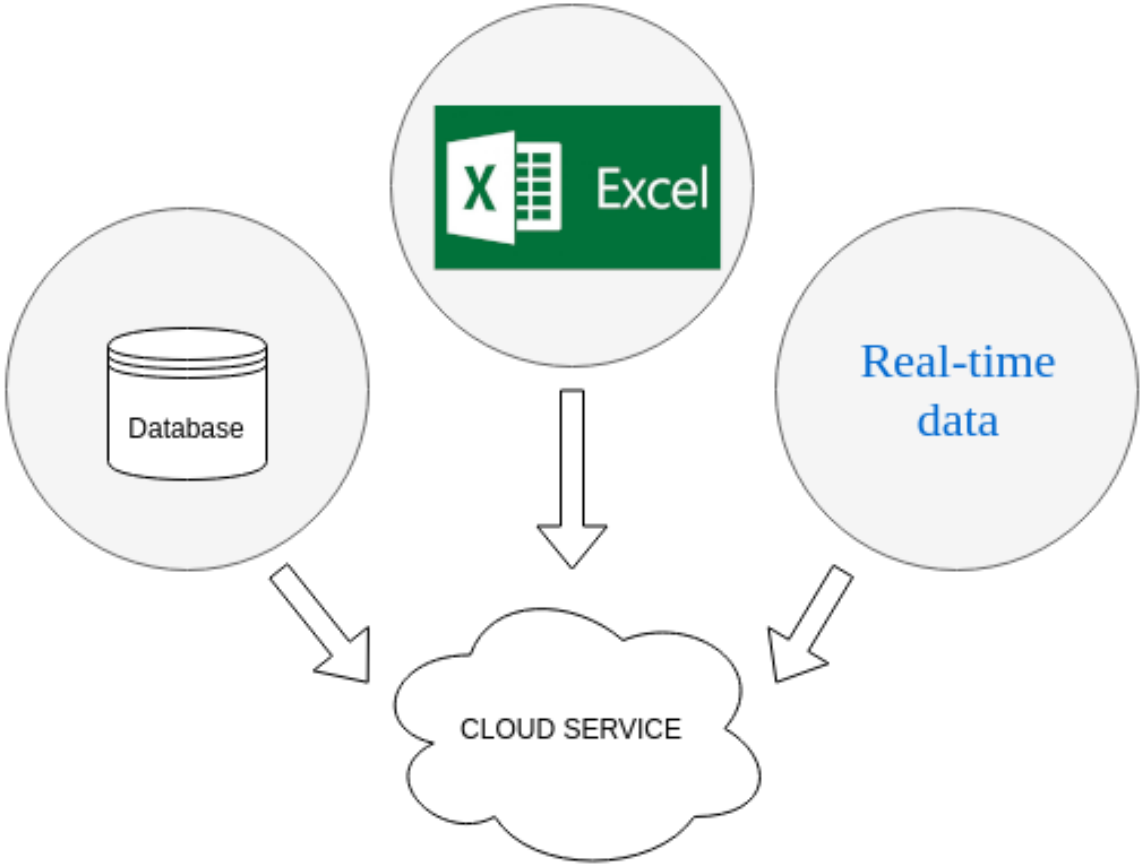
- Data collection, usually from diverse sources
- Data cleansing and preparation, including exploratory visualizations
- Choosing a machine learning model that suits our data
- Training the model with historical data (if we are talking about supervised learning)
- Mining the data for hidden or unknown patterns (if we are talking about unsupervised learning)
- Testing the accuracy of the model prediction
- Fine-tuning the model parameters or changing the model if the results are not satisfactory
- Visualizing the results
- Periodically re-training the model with new data


COLLECT
CLEANSE AND PREPARE
TRAIN & TEST
VISUALIZE
RAW DATA
DATABASE
MACHINE LEARNING MODEL
MODEL EVALUATION
DASHBOARDS
SUPERVISED LEARNING
DIVERSE DATA SOURCES
MISSING DATA
ERRONEOUS DATA
CREATE FEATURES
INTERACTIVE VISUALIZATIONS
FINE - TUNE OR CHANGE MODEL
COLLECT
CLEANSE AND PREPARE
DATA MINING
VISUALIZE
RAW DATA
DATABASE
MACHINE LEARNING MODEL
MODEL EVALUATION
DASHBOARDS
UNSUPERVISED LEARNING
DIVERSE DATA SOURCES
MISSING DATA
ERRONEOUS DATA
CREATE FEATURES
INTERACTIVE VISUALIZATIONS
FINE - TUNE OR CHANGE MODEL

Automatic data analysis flows

Data collection

- Once the different data sources (on-premise files and databases) are identified, the data can be periodically uploaded to a cloud storage service.
- This is usually done automatically by a process running periodically, with minimal human intervention.
- There are many different storage options available from the main cloud service providers.


Excel
Database
Real-time
data
CLOUD SERVICE

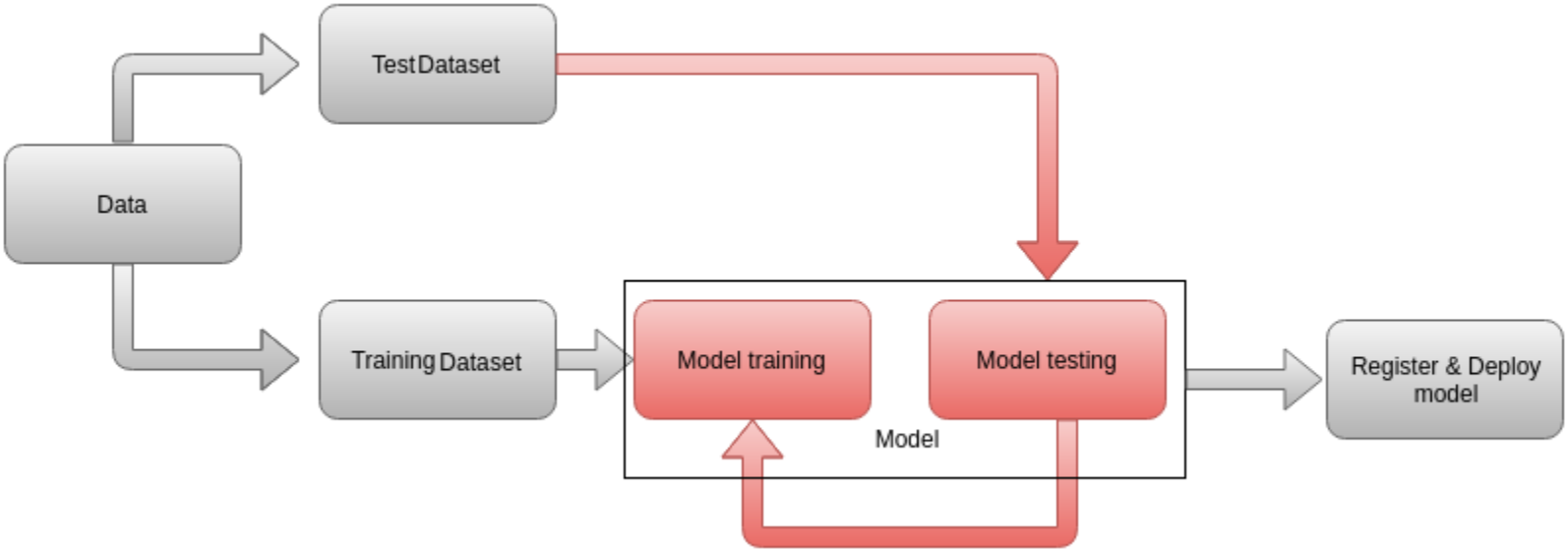
Automatic data analysis flows

Data preparation

- The full data cycle is shown in the following diagram:

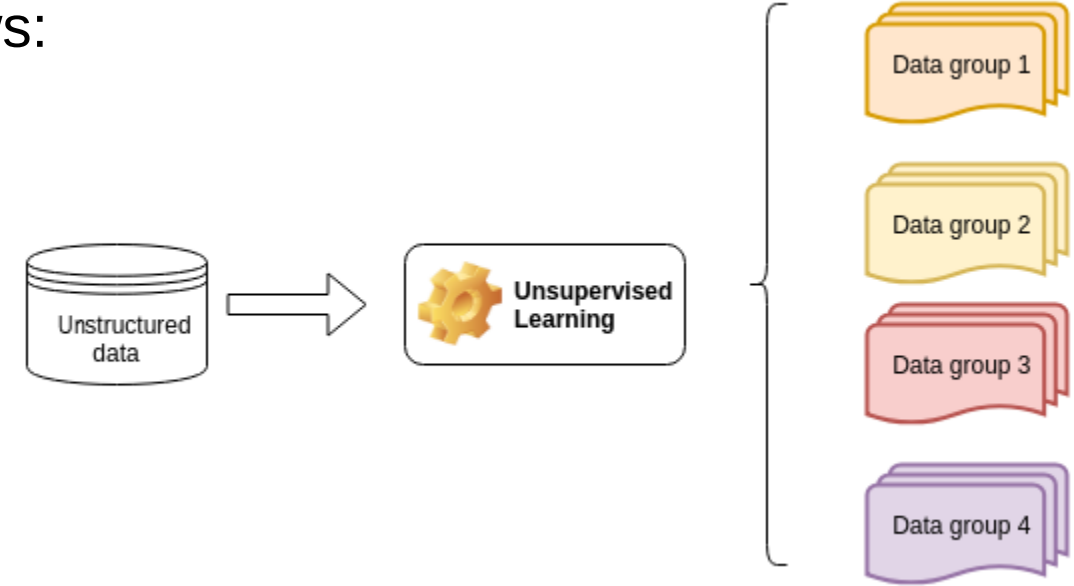
Model training

- The training cycle of a supervised machine learning model can be summarized as follows:



Unsupervised learning

- Whenever we are not sure of what we are going to find in data or we need to process a very large number of entries that would be impossible to manage manually, we use unsupervised machine learning. A general diagram could be as follows:



Automatic data analysis flows

Visualizations

- The last step in the data flow is visualization. When presenting our results to a non-technical audience, stressing the benefits of our analysis is of paramount importance to show the value of what we do.
- Interactive dashboards are the usual way of doing this, with advanced tools such as Tableau, Power BI, or QlikView. Some examples can be found at the following URL: https://www.clearpointstrategy.com/executive-dashboard-examples/.

Automatic data analysis flows

Re-training of machine learning models

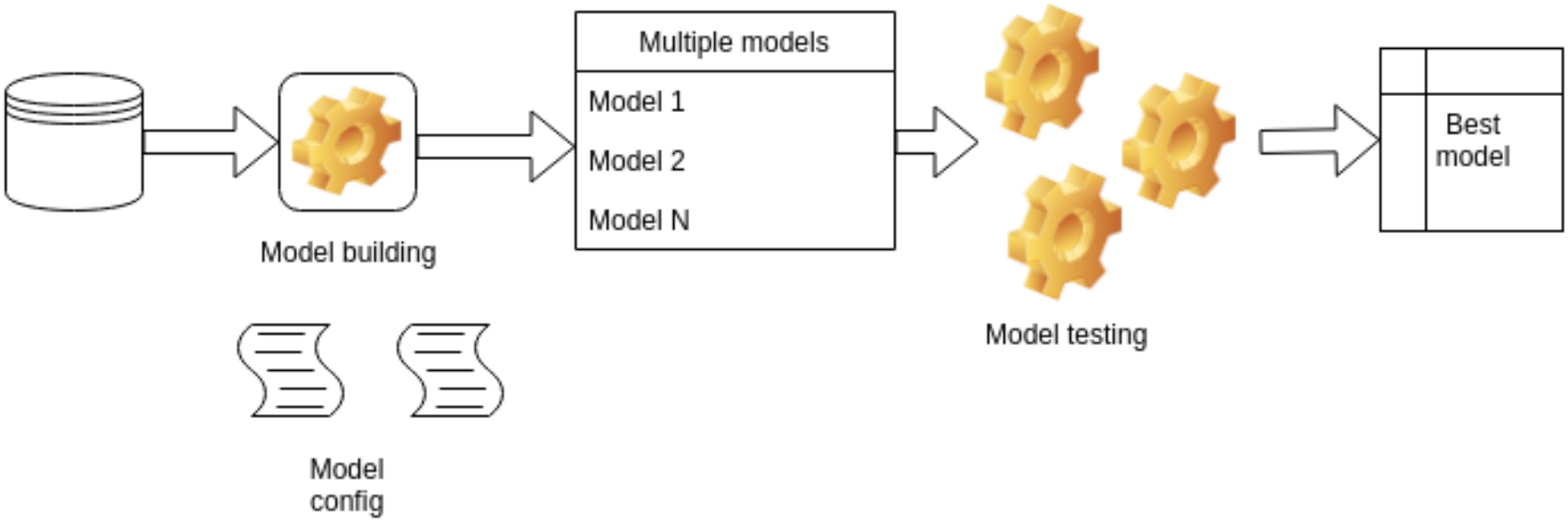
- Since new data is available all the time and business conditions change, machine learning models need periodic re-training.
- Cloud services offer ways of doing this with minimum intervention, without the need to rebuild any part of the data flow.
- You only need to load new data and specify that you are not building a new model but re-training an existing one. After finishing, the model will be available for use as usual.

Automated machine learning

There are several tasks that are crucial for the success of a machine learning model when applied to solve a given business problem, for example:

- Data pre-processing
- Feature engineering
- Model selection
- Optimization of the model hyperparameters
- Analysis of the model results

Automated machine learning



Automated machine learning

Following is the process for building of new model:

- Input data is pre-processed and used to build the best model features
- Based on some configurations done by the user, a given set of models is built and tested
- Models are evaluated and tested based on some criteria

Summary

- The last lesson of the course is thought of both as a summary of all lessons and as a window to what can be done beyond Excel and in the future.

- Automated data flows and machine learning model generation simplify analysts' work and speed up the decision-making process.

THANK YOU ☺