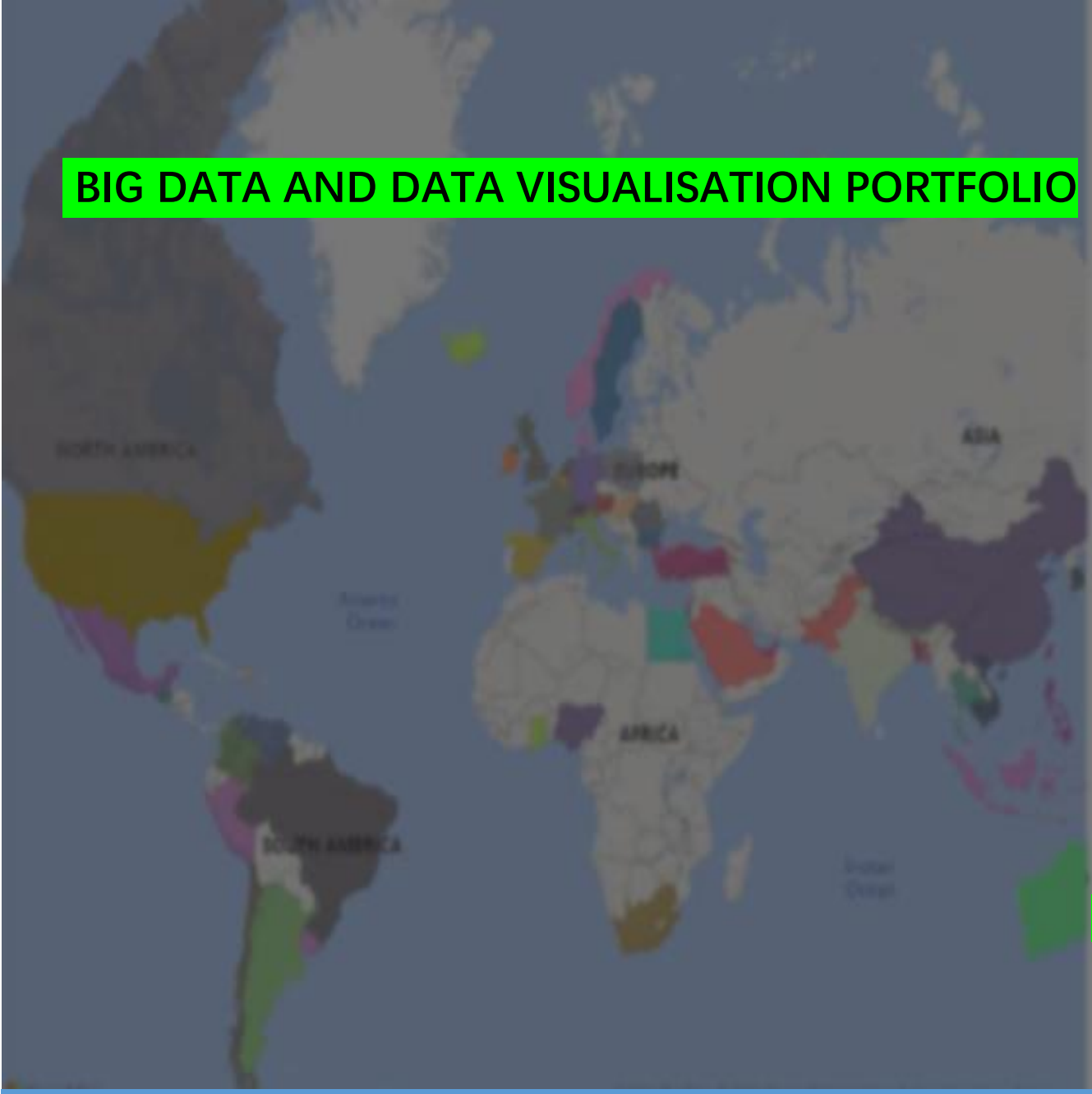
BIG DATA AND DATA VISUALISATION PORTFOLIO

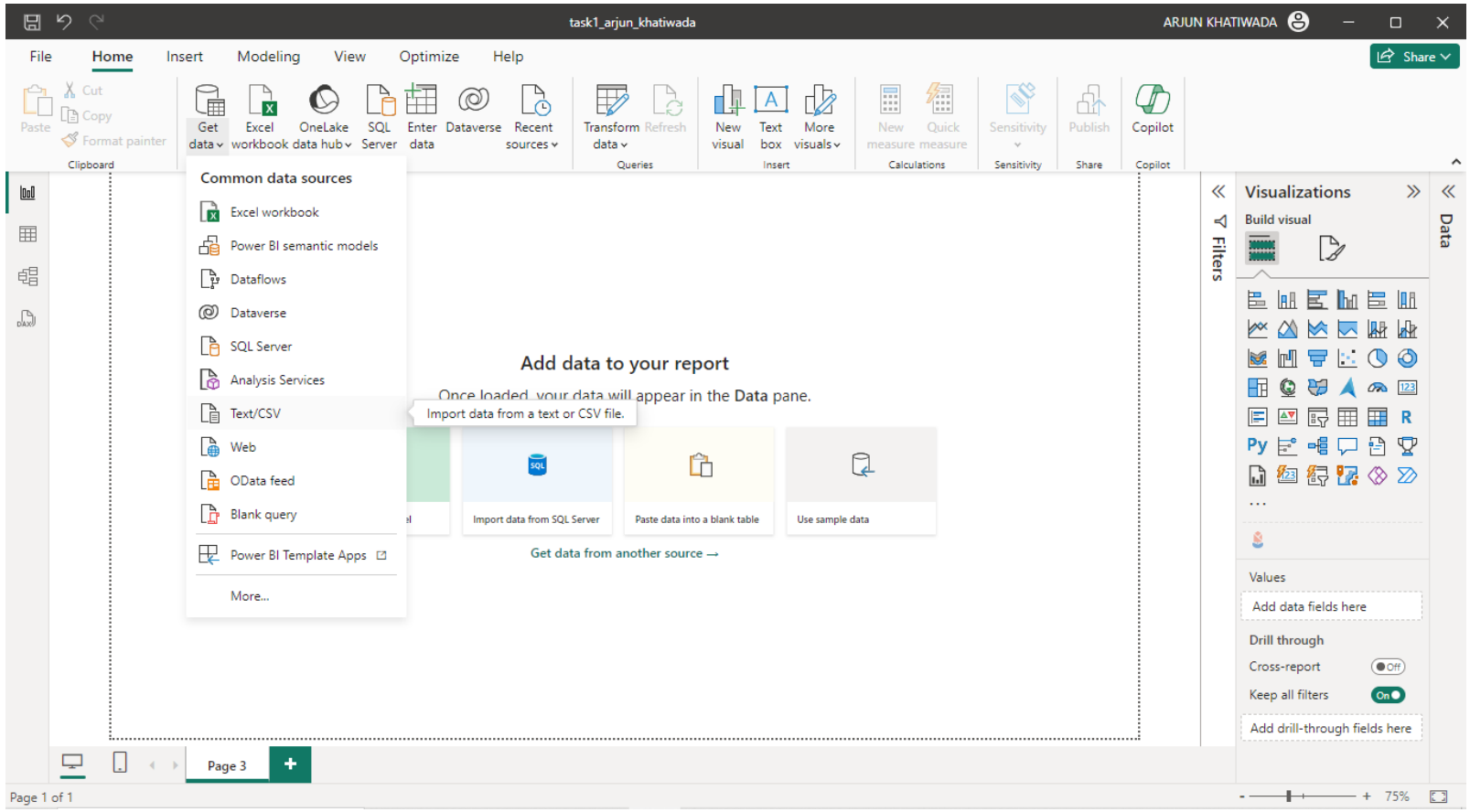


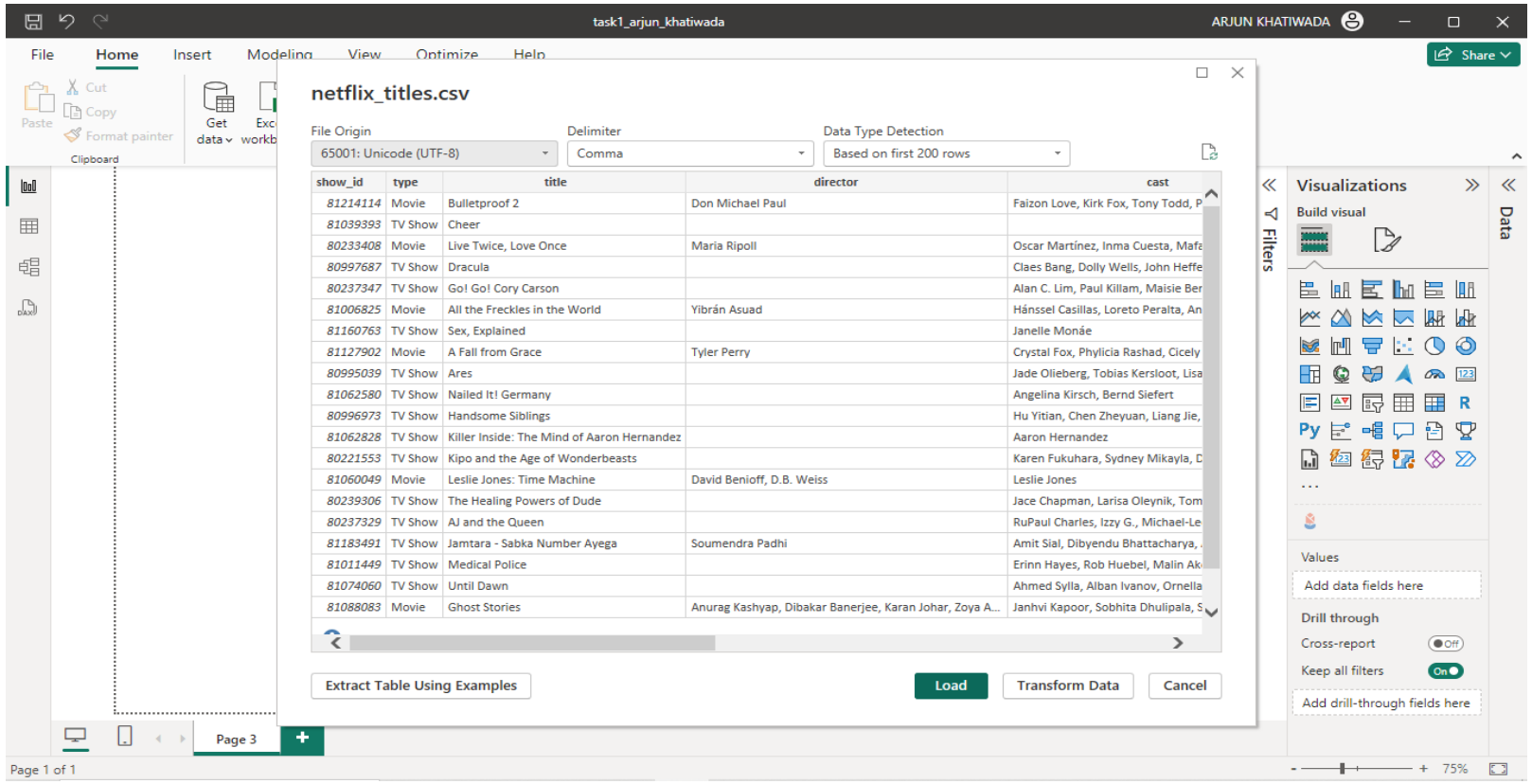
Submitted By - Arjun Khatiwada(S3098579)

Power BI - Assessment Sheet		
Student Name	Arjun Khatiwada	
Student ID	S3098579	

Task 1: Data Modelling & pre-processing (20 points total)

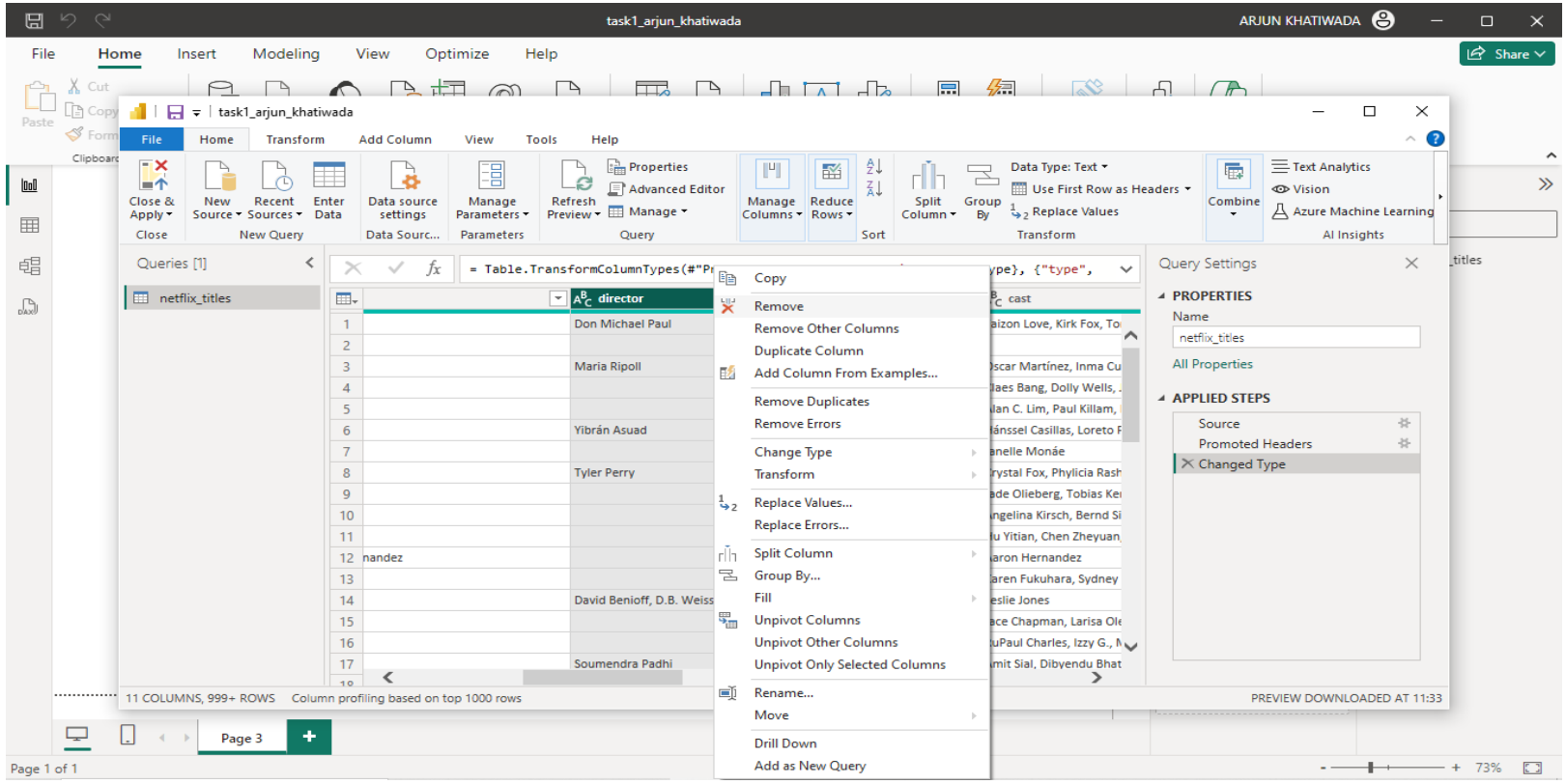
1) Import the data into Power BI

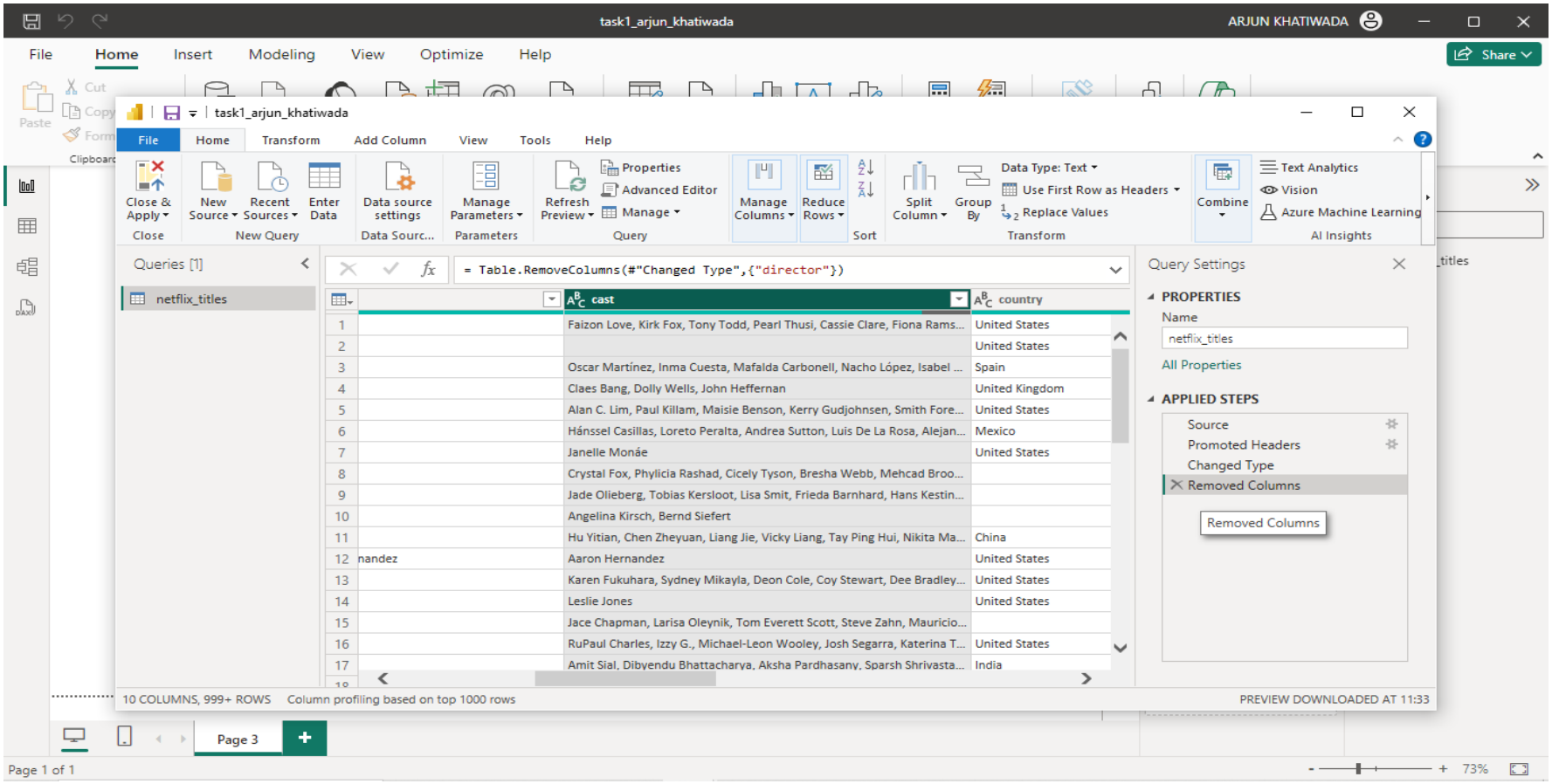




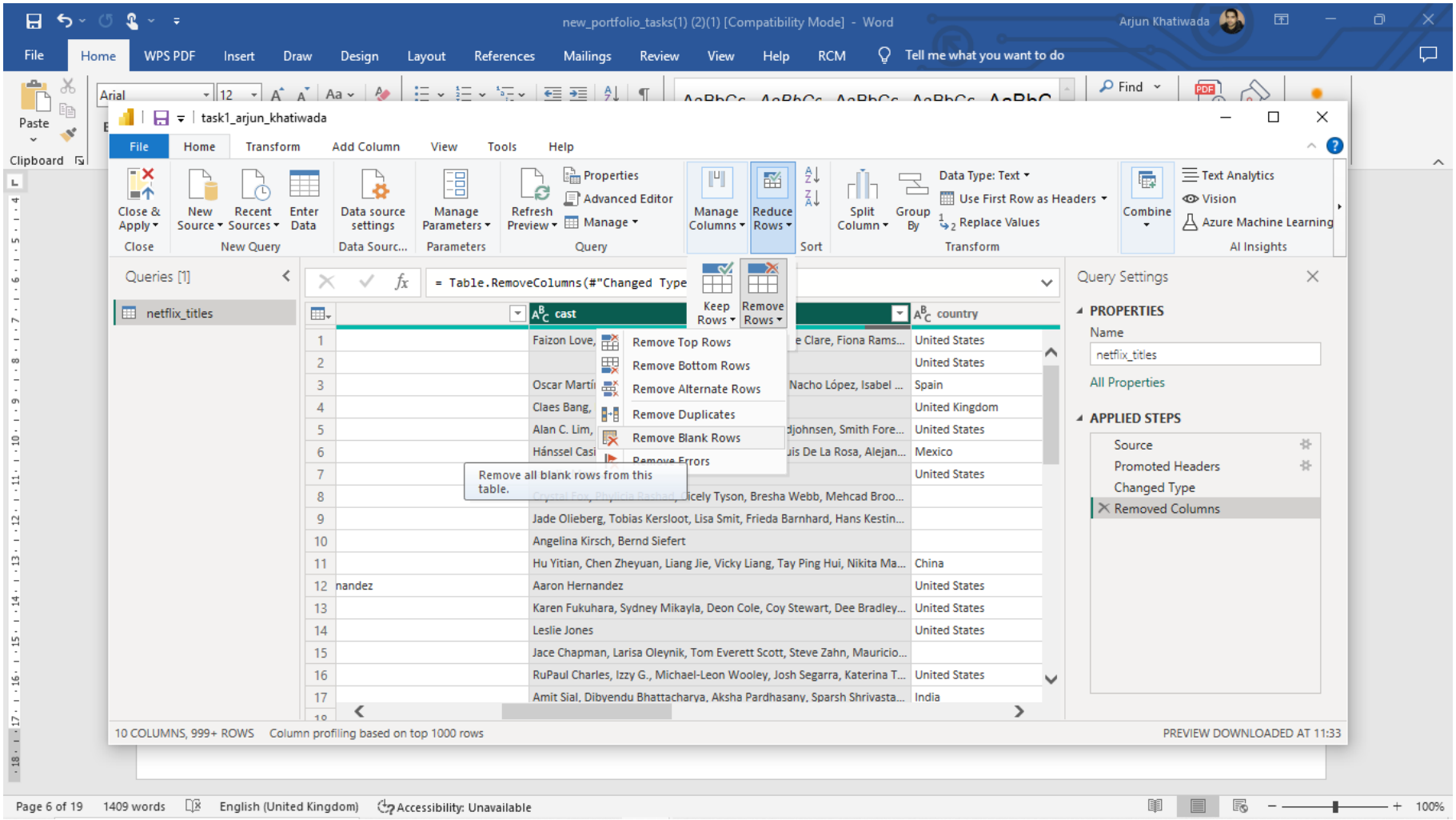
show_id	type	title	director	cast
81214114	Movie	Bulletproof 2	Don Michael Paul	Faizon Love, Kirk Fox, Tony Todd, P
81039393	TV Show	Cheer		
80233408	Movie	Live Twice, Love Once	Maria Ripoll	Oscar Martínez, Inma Cuesta, Mafa
80997687	TV Show	Dracula		Claes Bang, Dolly Wells, John Heffe
80237347	TV Show	Go! Go! Cory Carson		Alan C. Lim, Paul Killam, Maisie Ber
81006825	Movie	All the Freckles in the World	Yibrán Asuad	Hánssel Casillas, Loreto Peralta, An
81160763	TV Show	Sex, Explained		Janelle Monáe
81127902	Movie	A Fall from Grace	Tyler Perry	Crystal Fox, Phylicia Rashad, Cicely
80995039	TV Show	Ares		Jade Olieberg, Tobias Kersloot, Lisa
81062580	TV Show	Nailed It! Germany		Angelina Kirsch, Bernd Siefert
80996973	TV Show	Handsome Siblings		Hu Yitian, Chen Zheyuan, Liang Jie,
81062828	TV Show	Killer Inside: The Mind of Aaron Hernandez		Aaron Hernandez
80221553	TV Show	Kipo and the Age of Wonderbeasts		Karen Fukuhara, Sydney Mikayla, D
81060049	Movie	Leslie Jones: Time Machine	David Benioff, D.B. Weiss	Leslie Jones
80239306	TV Show	The Healing Powers of Dude		Jace Chapman, Larisa Oleynik, Tom
80237329	TV Show	AJ and the Queen		RuPaul Charles, Izzy G., Michael-Le
81183491	TV Show	Jamtara - Sabka Number Ayega	Soumendra Padhi	Amit Sial, Dibyendu Bhattacharya,
81011449	TV Show	Medical Police		Erinn Hayes, Rob Huebel, Malin Ak
81074060	TV Show	Until Dawn		Ahmed Sylla, Alban Ivanov, Ornella
81088083	Movie	Ghost Stories	Anurag Kashyap, Dibakar Banerjee, Karan Johar, Zoya A...	Janhvi Kapoor, Sobhita Dhulipala, S

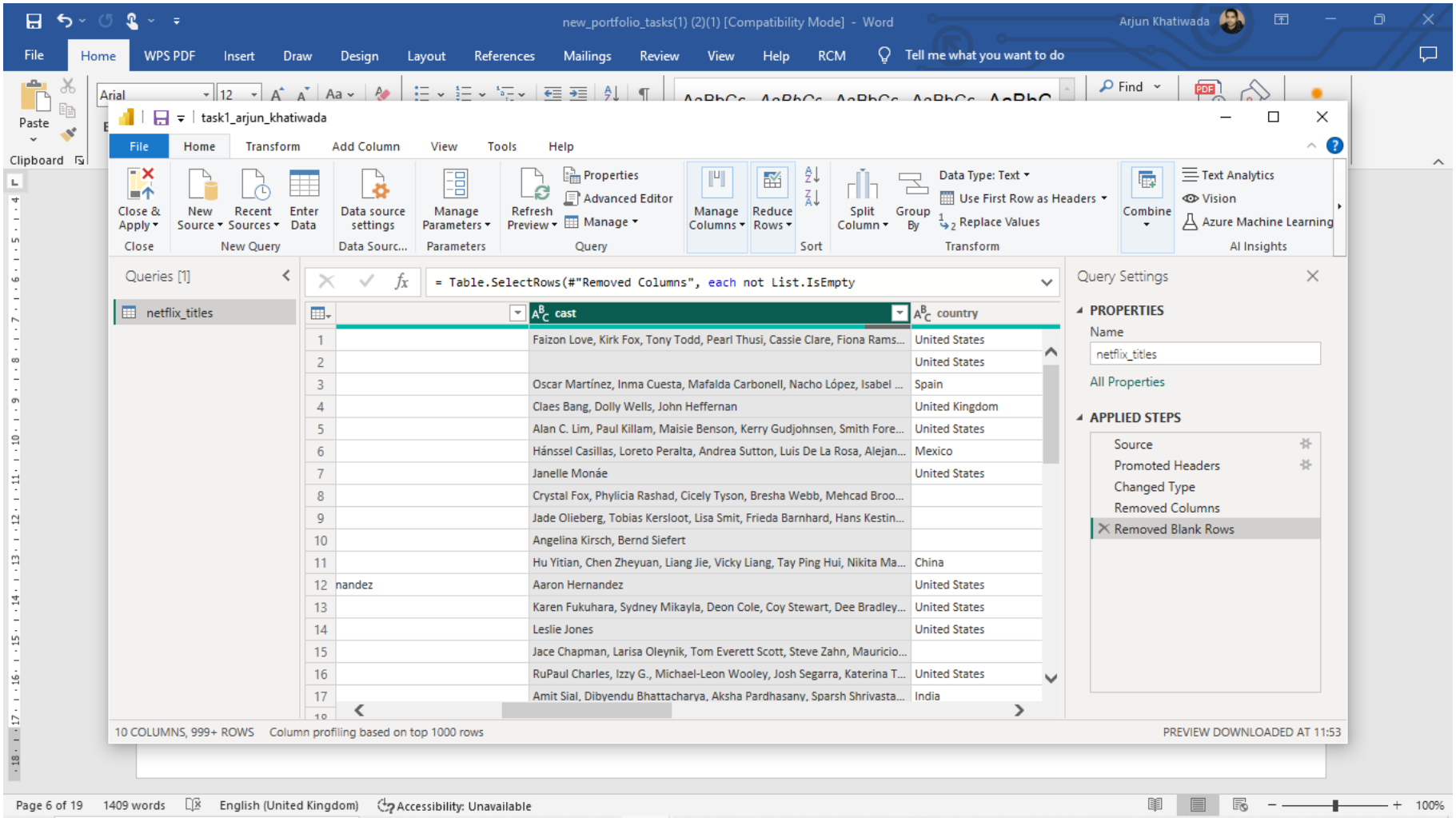
2) Deleting the column “director”



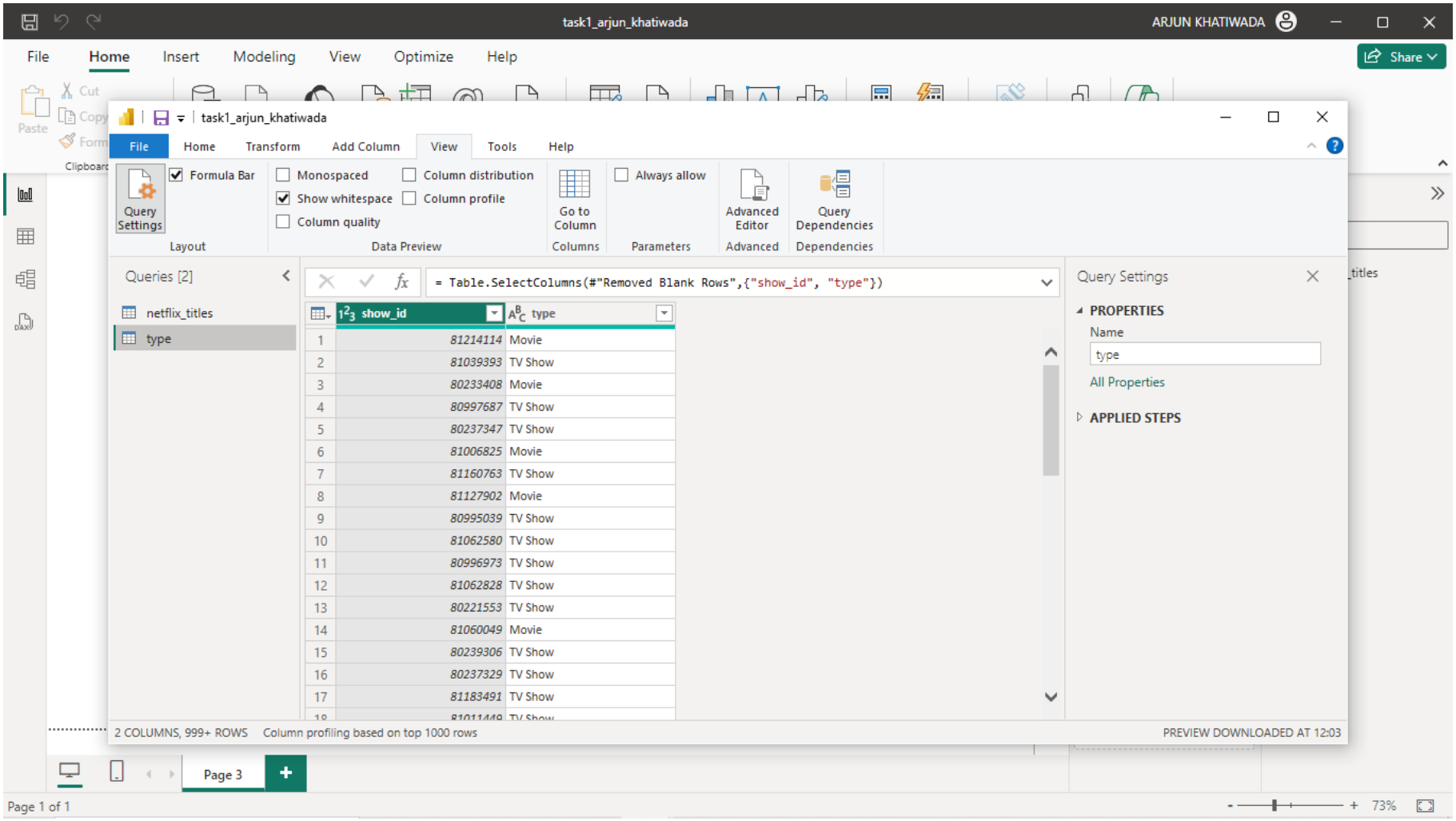


3) Removing all empty rows in ‘cast’ column

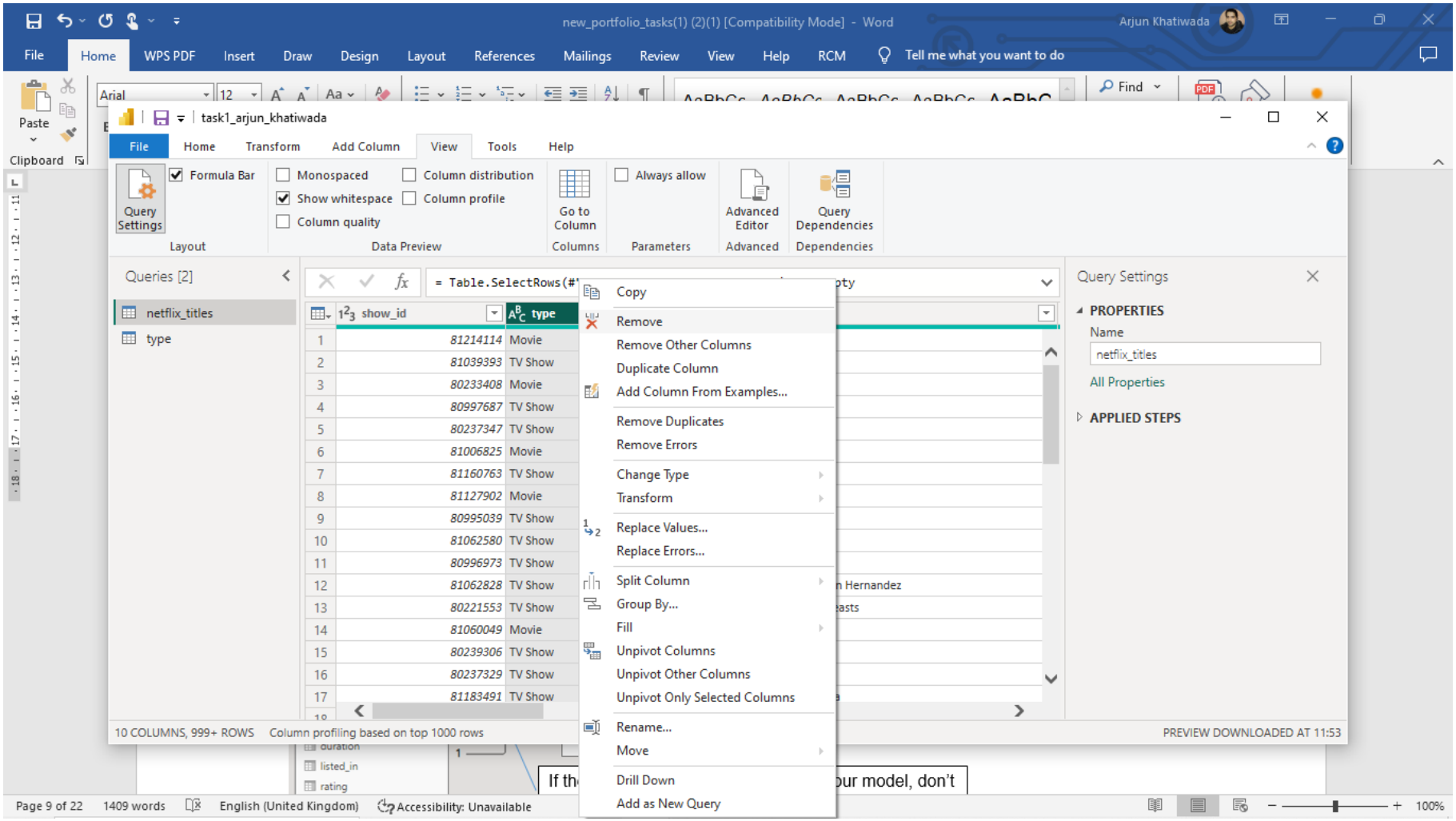

new_portfolio_tasks(1) (2)(1) [Compatibility Mode] - Word
Arjun Khatiwada
File
Home
WPS PDF
Insert
Draw
Design
Layout
References
Mailings
Review
View
Help
RCM
Tell me what you want to do
Paste
Clipboard
Find
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
Close
New Source
Recent Sources
Enter Data
New Query
Data source settings
Data Sourc...
Manage Parameters
Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Query
Manage Columns
Reduce Rows
Sort
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Transform
Combine
Text Analytics
Vision
Azure Machine Learning
AI Insights
Queries [1]
netflix_titles
= Table.RemoveColumns(#"Changed Type
Keep Rows
Remove Rows
Remove Top Rows
Remove Bottom Rows
Remove Alternate Rows
Remove Duplicates
Remove Blank Rows
Remove Errors
Remove all blank rows from this table.
cast
country
Faizon Love,
United States
United States
Oscar Martí
Nacho López, Isabel ...
Spain
Claes Bang,
United Kingdom
Alan C. Lim,
United States
Hánssel Casi
Mexico
United States
Jade Olieberg, Tobias Kersloot, Lisa Smit, Frieda Barnhard, Hans Kestin...
Angelina Kirsch, Bernd Siefert
Hu Yitian, Chen Zheyuan, Liang Jie, Vicky Liang, Tay Ping Hui, Nikita Ma...
China
Aaron Hernandez
United States
Karen Fukuhara, Sydney Mikayla, Deon Cole, Coy Stewart, Dee Bradley...
United States
Leslie Jones
United States
Jace Chapman, Larisa Oleynik, Tom Everett Scott, Steve Zahn, Mauricio...
RuPaul Charles, Izzy G., Michael-Leon Wooley, Josh Segarra, Katerina T...
United States
Amit Sial, Dibyendu Bhattacharya, Aksha Pardhasany, Sparsh Shrivasta...
India
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Removed Columns
10 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 11:33
Page 6 of 19
1409 words
English (United Kingdom)
Accessibility: Unavailable
100%

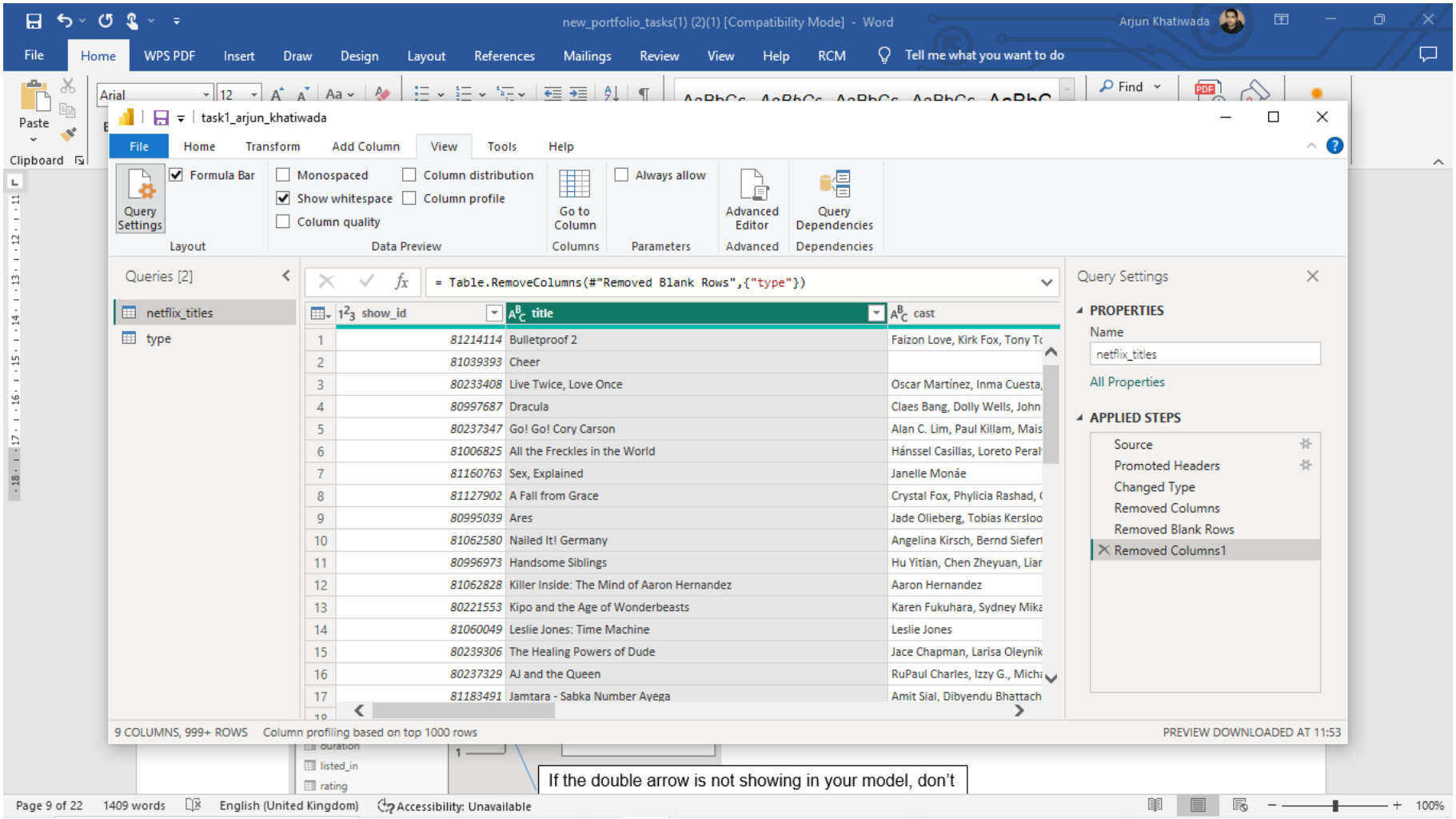


4) Creating a new table ‘type’ having columns ‘show_id’ and the ‘type’

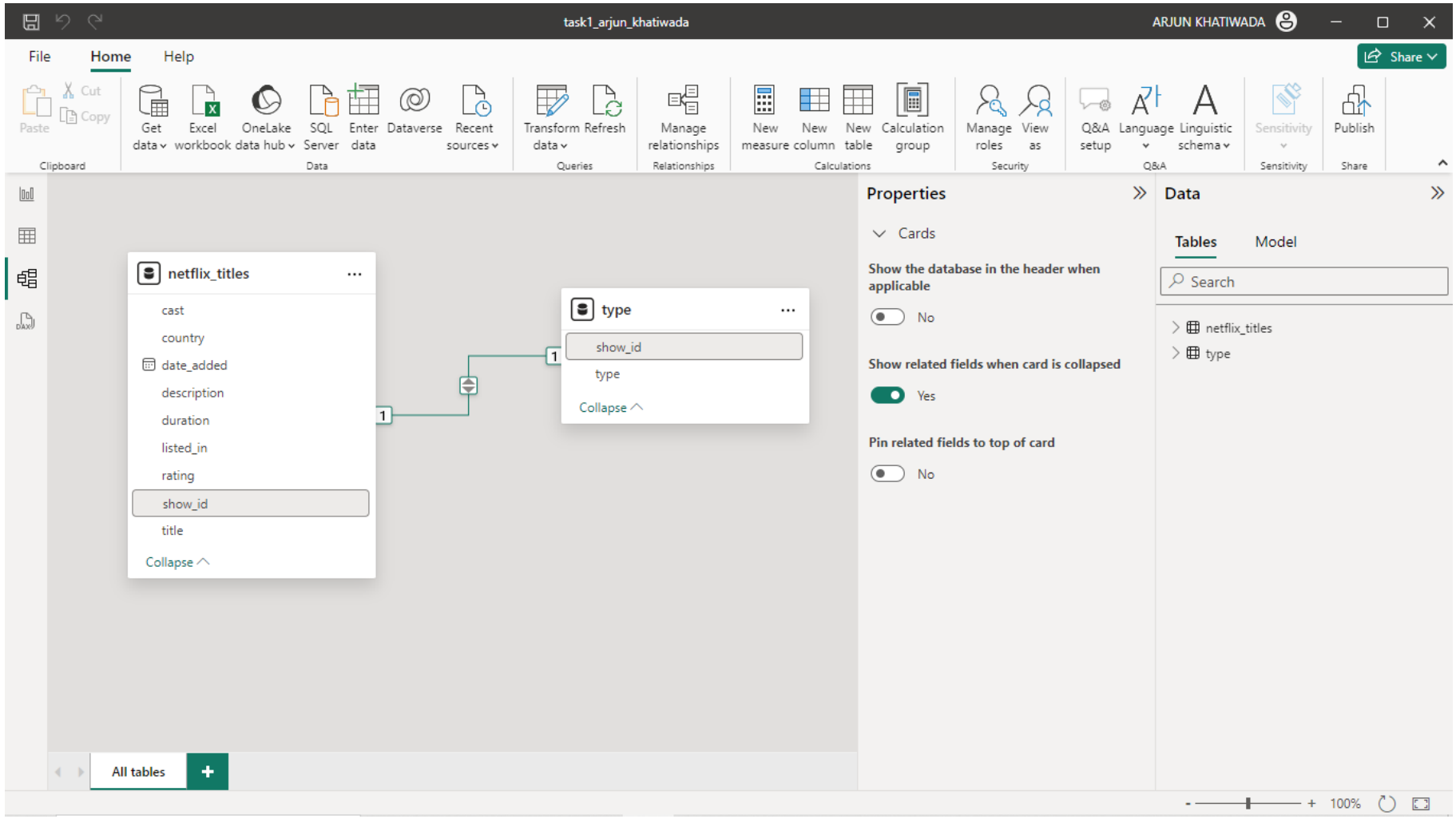


5) Removing the column ‘type’ from table Netflix_titles

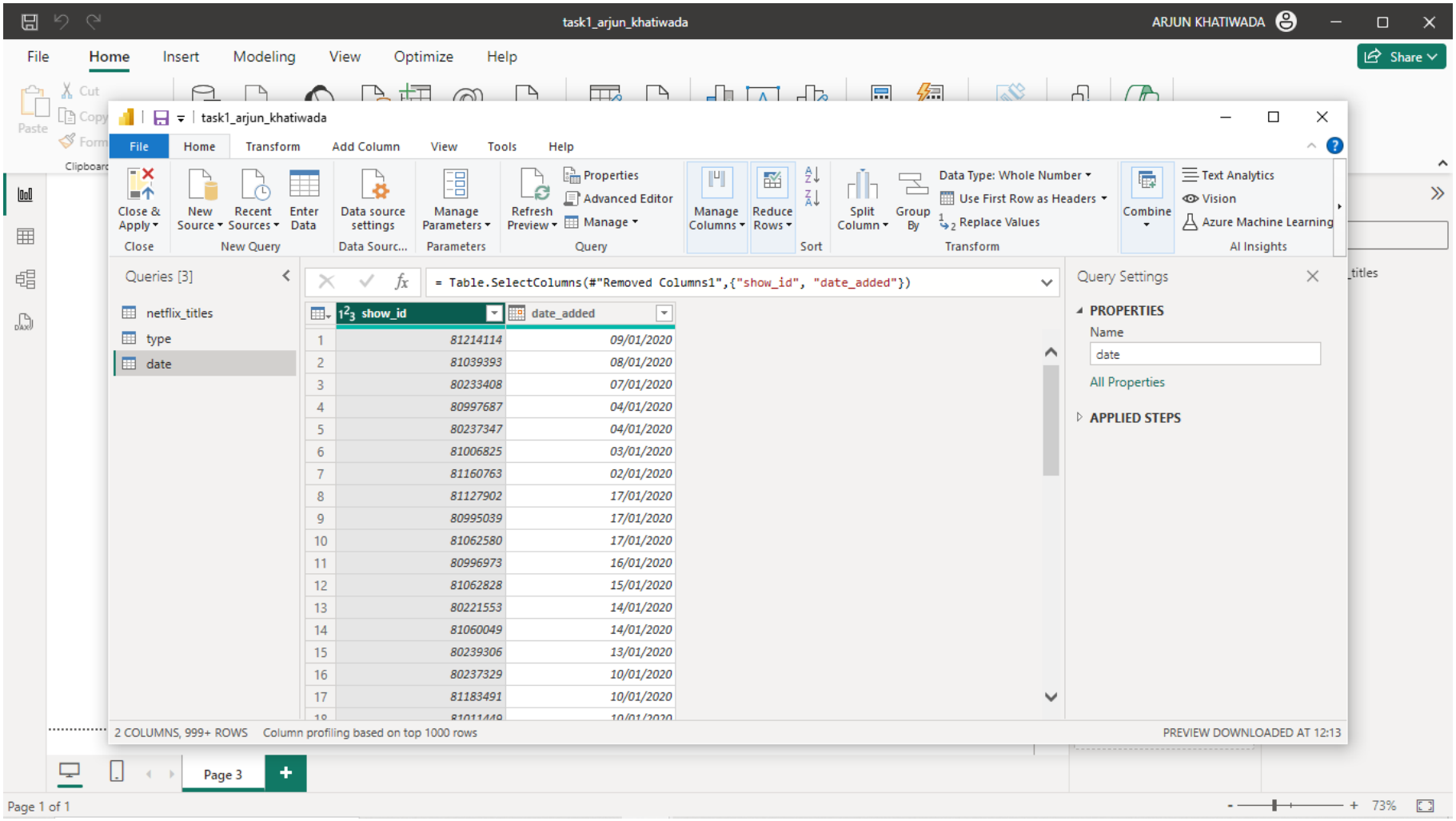

new_portfolio_tasks(1) (2)(1) [Compatibility Mode] - Word
Arjun Khatiwada
File
Home
WPS PDF
Insert
Draw
Design
Layout
References
Mailings
Review
View
Help
RCM
Tell me what you want to do
Paste
Clipboard
Find
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Query Settings
Formula Bar
Layout
Monospaced
Show whitespace
Column quality
Column distribution
Column profile
Data Preview
Go to Column
Columns
Always allow
Parameters
Advanced Editor
Advanced
Query Dependencies
Dependencies
Queries [2]
netflix_titles
type
= Table.SelectRows(#
show_id
type
1 81214114 Movie
2 81039393 TV Show
3 80233408 Movie
4 80997687 TV Show
5 80237347 TV Show
6 81006825 Movie
7 81160763 TV Show
8 81127902 Movie
9 80995039 TV Show
10 81062580 TV Show
11 80996973 TV Show
12 81062828 TV Show
13 80221553 TV Show
14 81060049 Movie
15 80239306 TV Show
16 80237329 TV Show
17 81183491 TV Show
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Split Column
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
10 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 11:53
Page 9 of 22
1409 words
English (United Kingdom)
Accessibility: Unavailable
100%



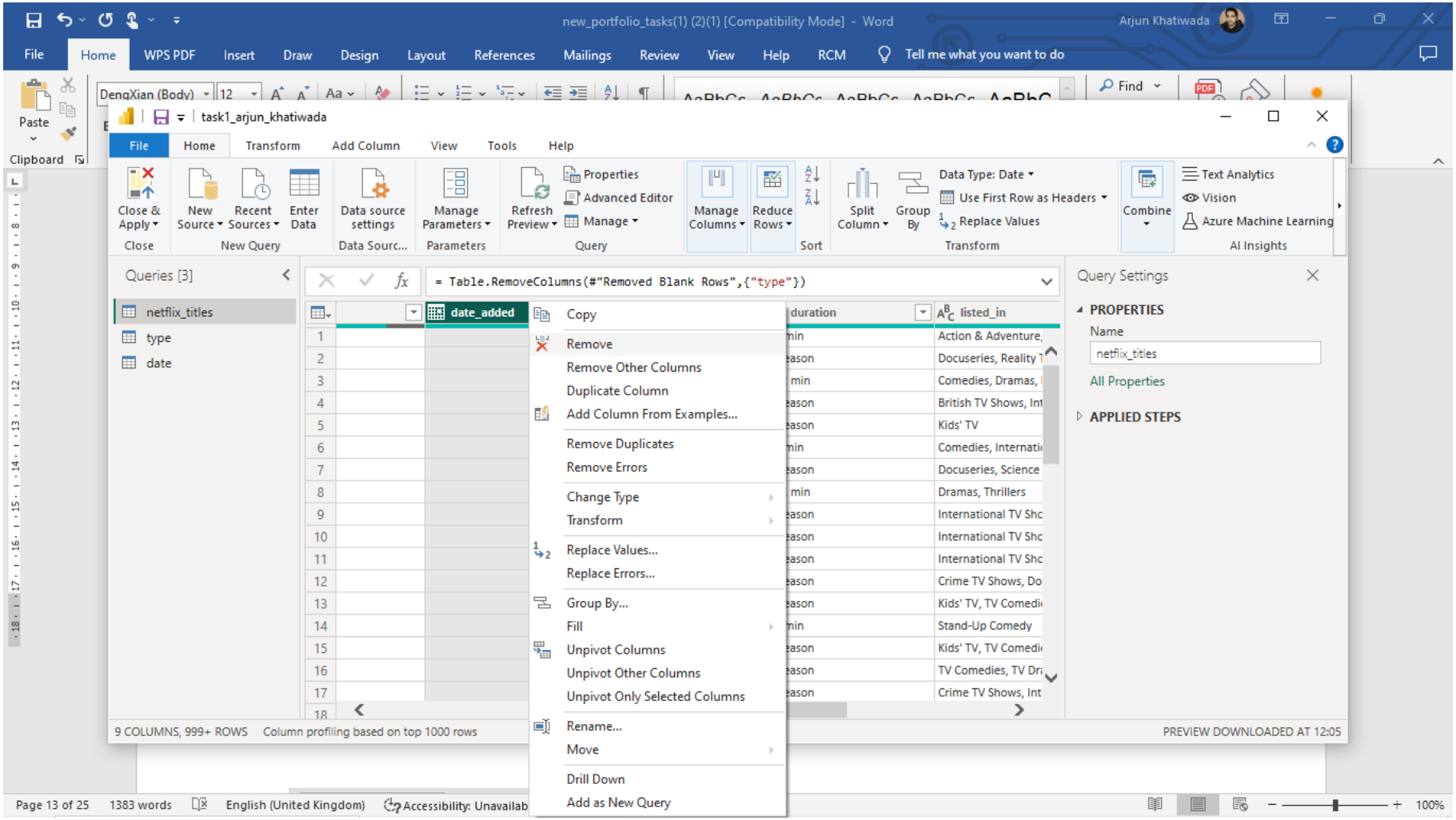
6) Using Close and apply. There is 1 to 1 relationship between two tables as shown below in the model view

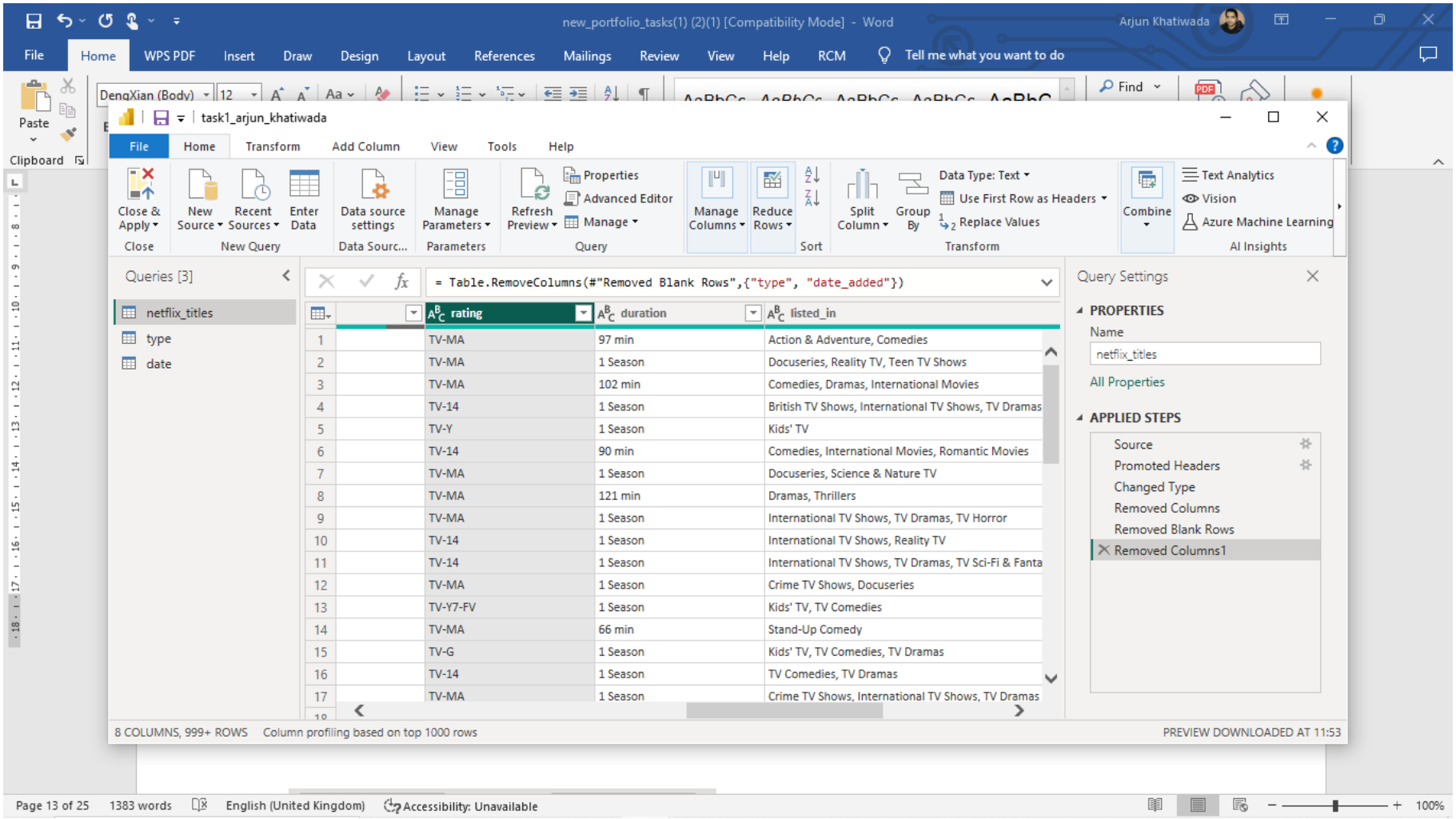

task1_arjun_khatiwada
ARJUN KHATIWADA
File
Home
Help
Share
Cut
Copy
Paste
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
Manage relationships
Relationships
New measure
New column
New table
Calculation group
Calculations
Manage roles
View as
Security
Q&A setup
Language
Linguistic schema
Q&A
Sensitivity
Publish
Share
netflix_titles
cast
country
date_added
description
duration
listed_in
rating
show_id
title
Collapse
type
show_id
type
Collapse
1
1
All tables
Properties
Cards
Show the database in the header when applicable
No
Show related fields when card is collapsed
Yes
Pin related fields to top of card
No
Data
Tables
Model
Search
netflix_titles
type
100%

7) Creating table “date” having columns show_id and date_added.

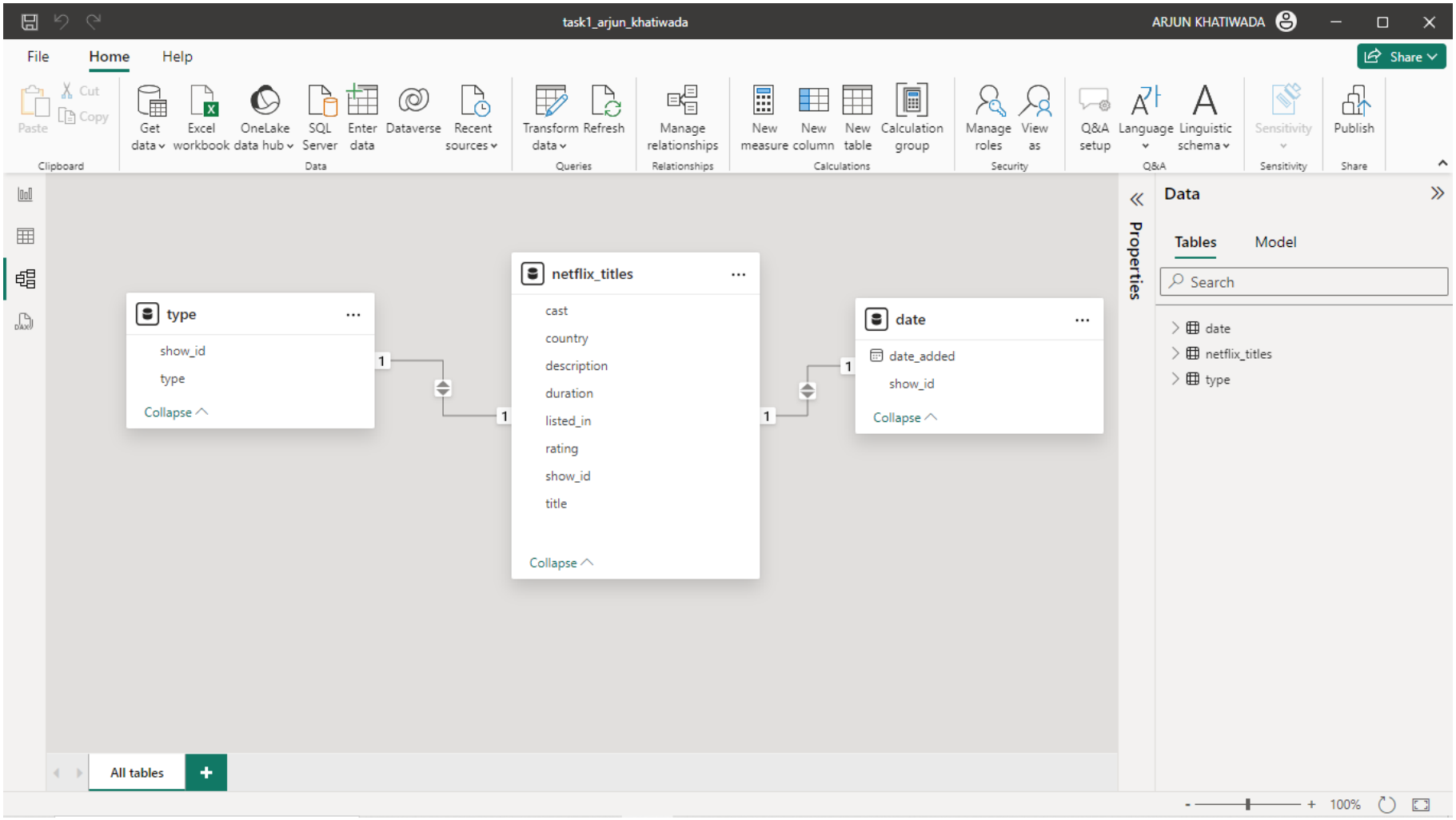


8) Removing the column date_added from the table Netflix_titles.


new_portfolio_tasks(1) (2)(1) [Compatibility Mode] - Word
Arjun Khatiwada
File
Home
WPS PDF
Insert
Draw
Design
Layout
References
Mailings
Review
View
Help
RCM
Tell me what you want to do
Paste
Clipboard
Find
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Manage Columns
Reduce Rows
Split Column
Group By
Data Type: Date
Use First Row as Headers
Replace Values
Combine
Text Analytics
Vision
Azure Machine Learning
Close
New Query
Data Sourc...
Parameters
Query
Sort
Transform
AI Insights
Queries [3]
netflix_titles
type
date
= Table.RemoveColumns(#"Removed Blank Rows",{"type"})
date_added
duration
listed_in
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
9 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 12:05
Page 13 of 25
1383 words
English (United Kingdom)
Accessibility: Unavailab
100%



9) Using Close & Apply and checking model view to see three tables like below:



10) Removing all the rows having type “TV Shows” from the given dataset

task1_arjun_khatiwada

ARJUN KHATIWADA

File Home Insert Modeling View Optimize Help

Share

Cut Copy Paste Clipboard

task1_arjun_khatiwada

File Home Transform Add Column View Tools Help

Close & Apply | New Source | Recent Sources | Enter Data | Data source settings | Manage Parameters | Refresh Preview | Properties | Advanced Editor | Manage | Manage Columns | Reduce Rows | Split Column | Group By | Data Type: Whole Number | Use First Row as Headers | Replace Values | Combine | Text Analytics | Vision | Azure Machine Learning

Close | New Query | Data Sourc... | Parameters | Query | Sort | Transform | AI Insights

Queries [3]

- netflix_titles
- type
- date

= Table.SelectRows(#"Removed Other Columns", each ([type] = "Movie"))

	show_id	type
1	81214114	Movie
2	80233408	Movie
3	81006825	Movie
4	81127902	Movie
5	81060049	Movie
6	81088083	Movie
7	81154455	Movie
8	81132437	Movie
9	81176188	Movie
10	81078908	Movie
11	81155784	Movie
12	81054495	Movie
13	81053893	Movie
14	80200087	Movie
15	81053892	Movie
16	80231903	Movie
17	81171121	Movie
18	81177238	Movie

Query Settings

PROPERTIES

Name: type

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Removed Blank Rows
- Removed Other Columns
- Filtered Rows

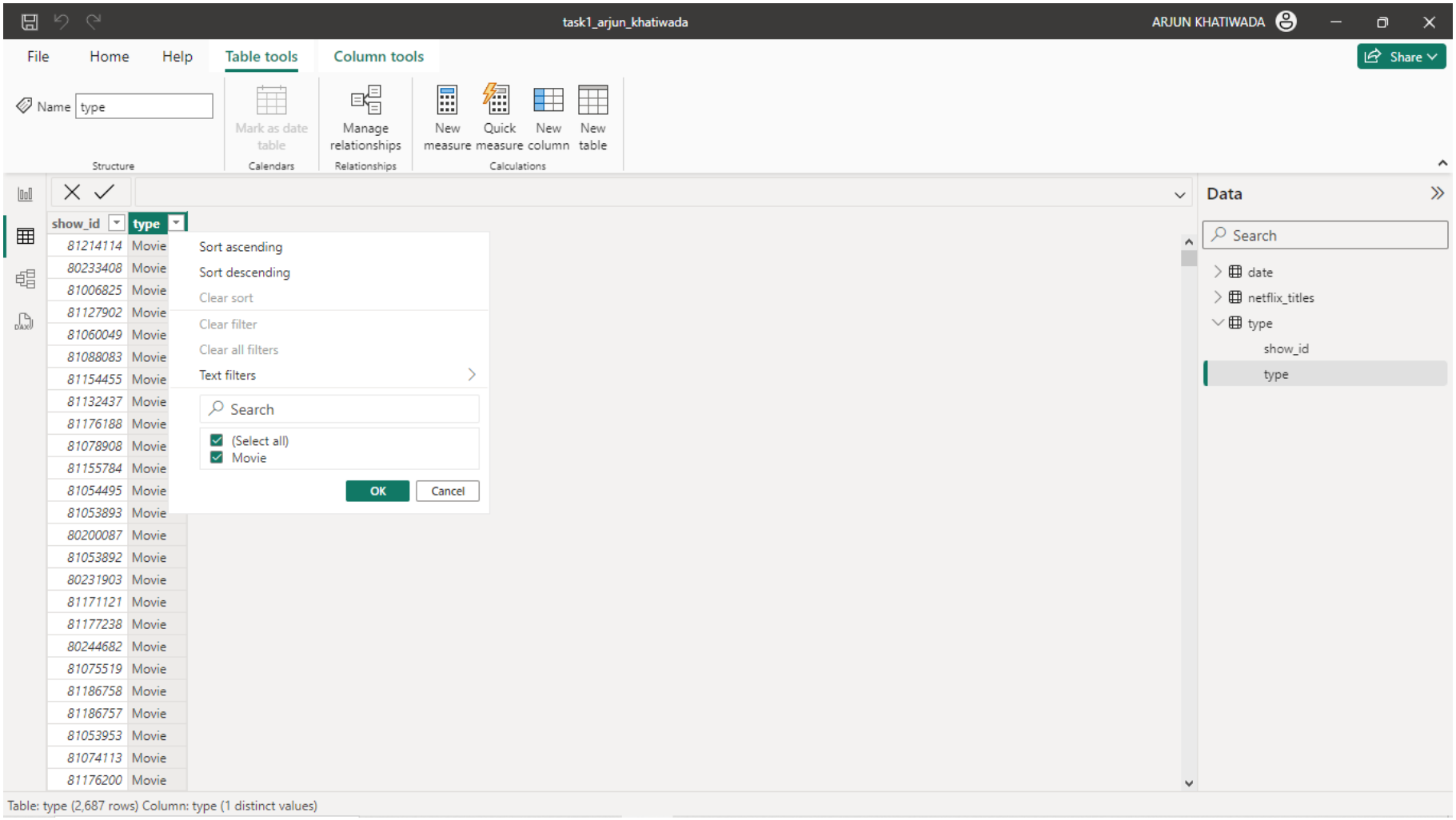
2 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 12:24

Page 3

Page 1 of 1

73%



11) Removing rows having seasons in duration column

task1_arjun_khatiwada

ARJUN KHATIWADA

File Home Insert Modeling View Optimize Help

Share

task1_arjun_khatiwada

File Home Transform Add Column View Tools Help

Close & Apply | New Source | Recent Sources | Enter Data | Data source settings | Manage Parameters | Refresh Preview | Properties | Advanced Editor | Manage | Manage Columns | Reduce Rows | Split Column | Group By | Data Type: Text | Use First Row as Headers | Replace Values | Combine | Text Analytics | Vision | Azure Machine Learning

Close | New Query | Data Sourc... | Parameters | Query | Sort | Transform | AI Insights

Queries [3]

- netflix_titles
- type
- date

```
= Table.SelectRows(#"Removed Columns1", each ([duration] <> "1 Season" and
    [duration] <> "11 Seasons" and [duration] <> "12 Seasons" and [duration] <> "13
    Seasons" and [duration] <> "14 Seasons" and [duration] <> "15 Seasons" and
    [duration] <> "2 Seasons" and [duration] <> "3 Seasons" and [duration] <> "4
    Seasons" and [duration] <> "5 Seasons" and [duration] <> "6 Seasons" and
```

	duration	listed_in	description
1	97 min	Action & Adventure, Comedies	A special agent ab
2	102 min	Comedies, Dramas, International Movies	When Emilio (Osc
3	90 min	Comedies, International Movies, Romantic Movies	Thirteen-year-old
4	121 min	Dramas, Thrillers	When gentle, law-
5	66 min	Stand-Up Comedy	From trying to sed
6	145 min	Horror Movies, International Movies, Thrillers	The directors of E
7	125 min	Dramas, International Movies, Thrillers	The grim realities
8	100 min	Comedies, International Movies, Romantic Movies	Bogged down with
9	10 min	Documentaries	Barack and Miche
10	108 min	Dramas, Romantic Movies	A teenage equestr
11	93 min	Comedies, Dramas, International Movies	Rushing to pay off
12	64 min	Stand-Up Comedy	Comedian Mo Gill
13	45 min	Documentaries, International Movies	An in-depth look a
14	89 min	International Movies, Thrillers	A Copenhagen pol
15	90 min	Dramas, International Movies	A clumsy but chee
16	116 min	Dramas, Sci-Fi & Fantasy, Thrillers	A Philadelphia det
17			

Query Settings

PROPERTIES

Name: netflix_titles

All Properties

APPLIED STEPS

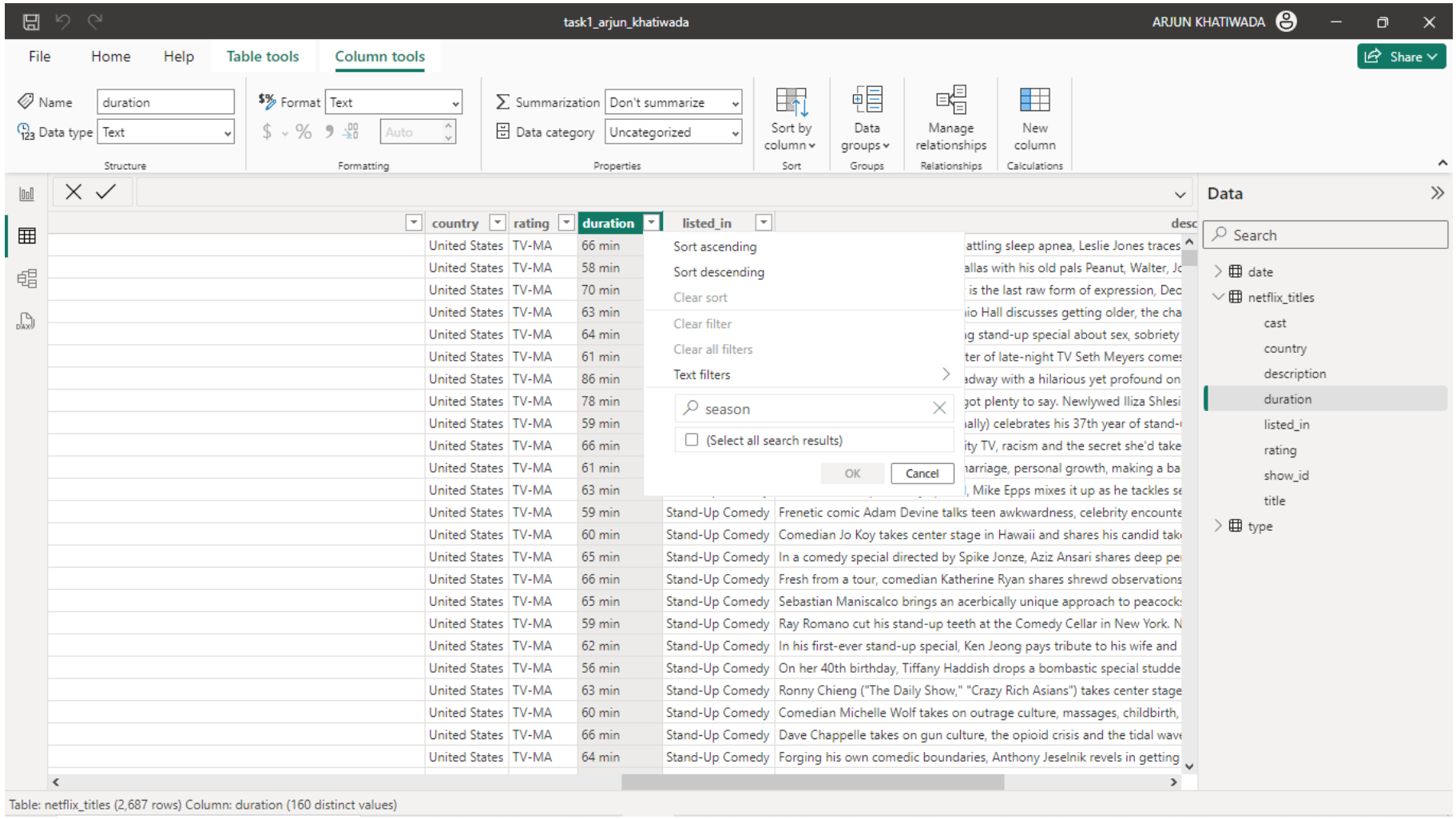
- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Removed Blank Rows
- Removed Columns1
- Filtered Rows

8 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

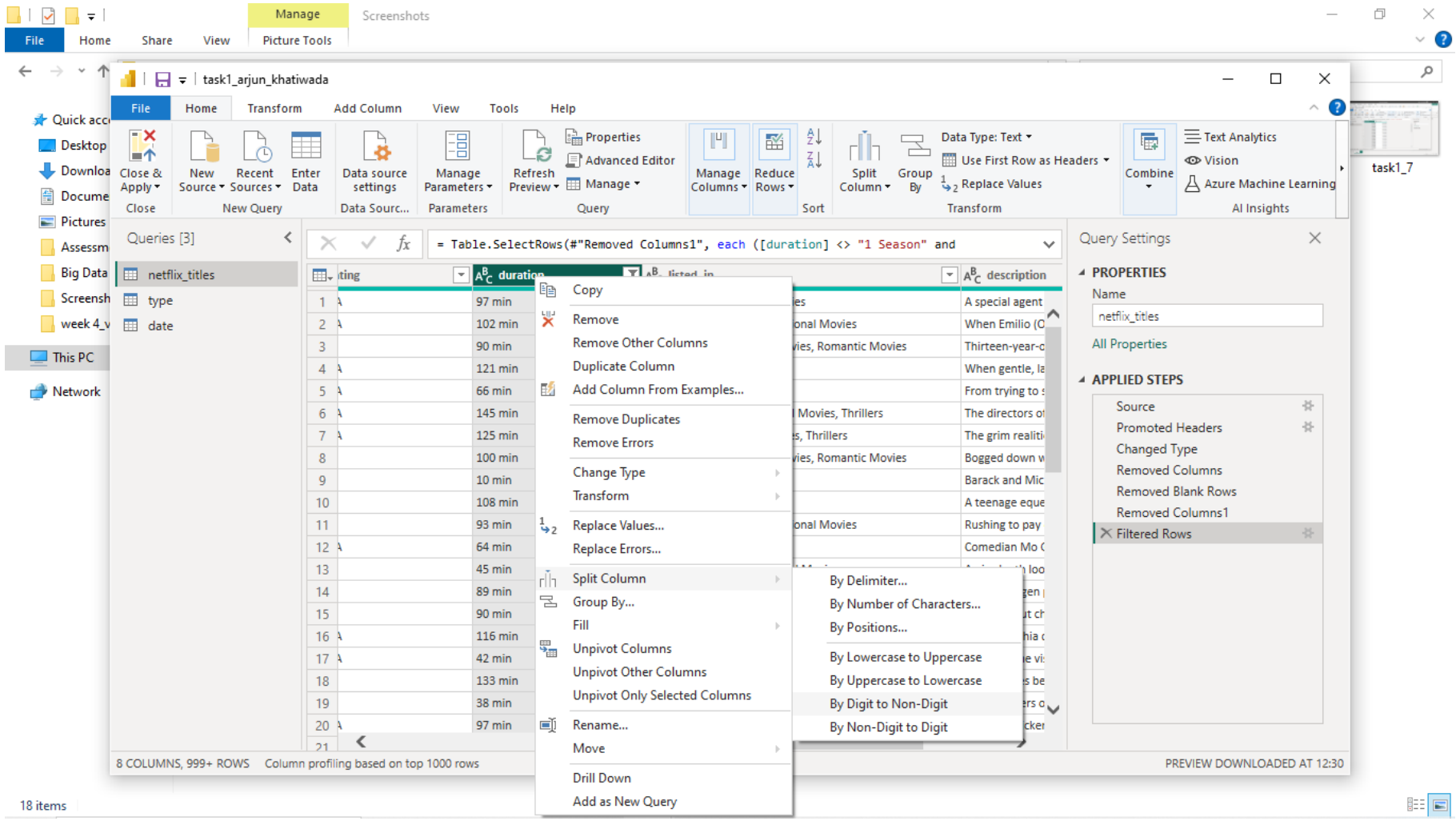
PREVIEW DOWNLOADED AT 12:30

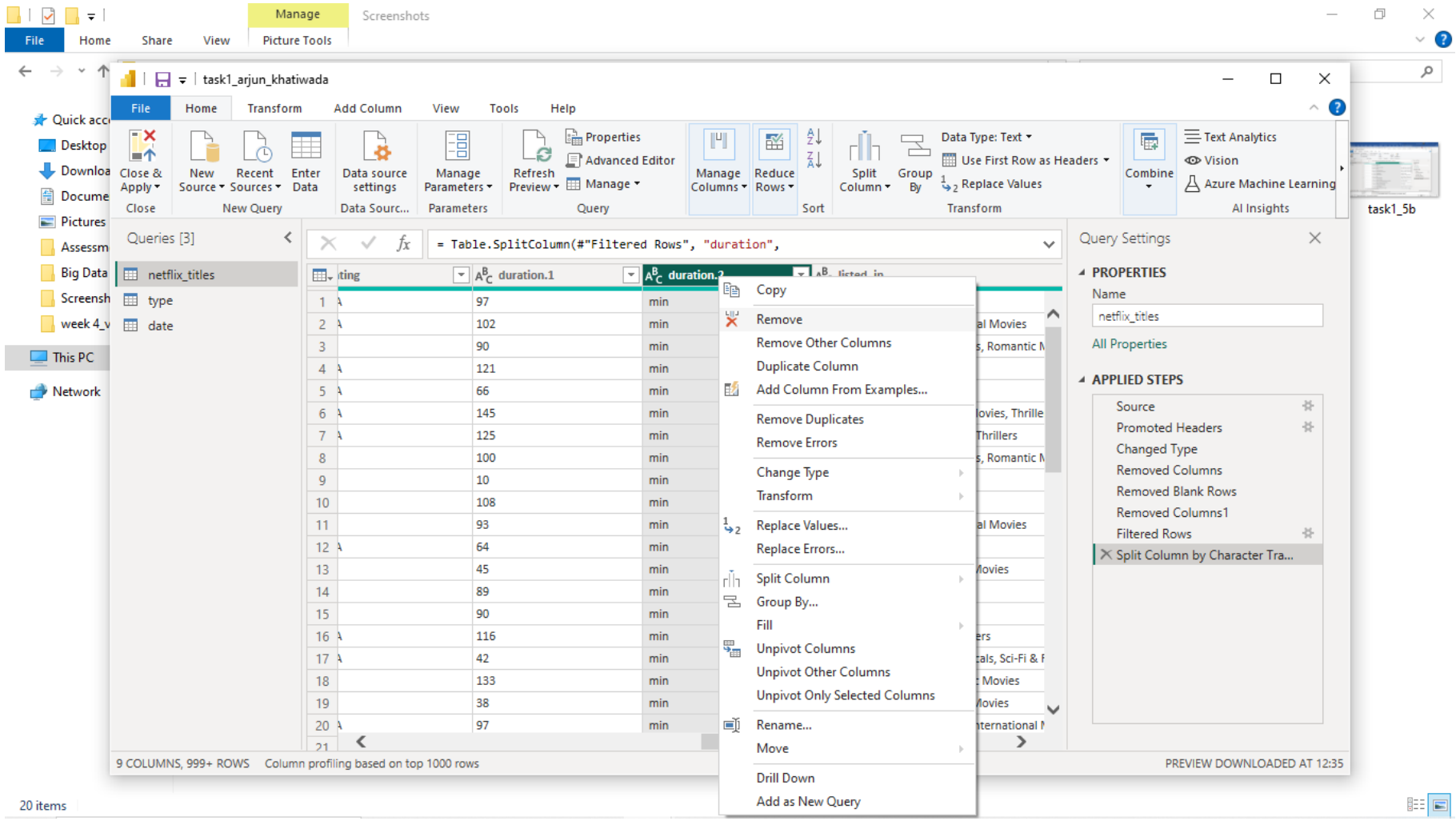
Page 1 of 1

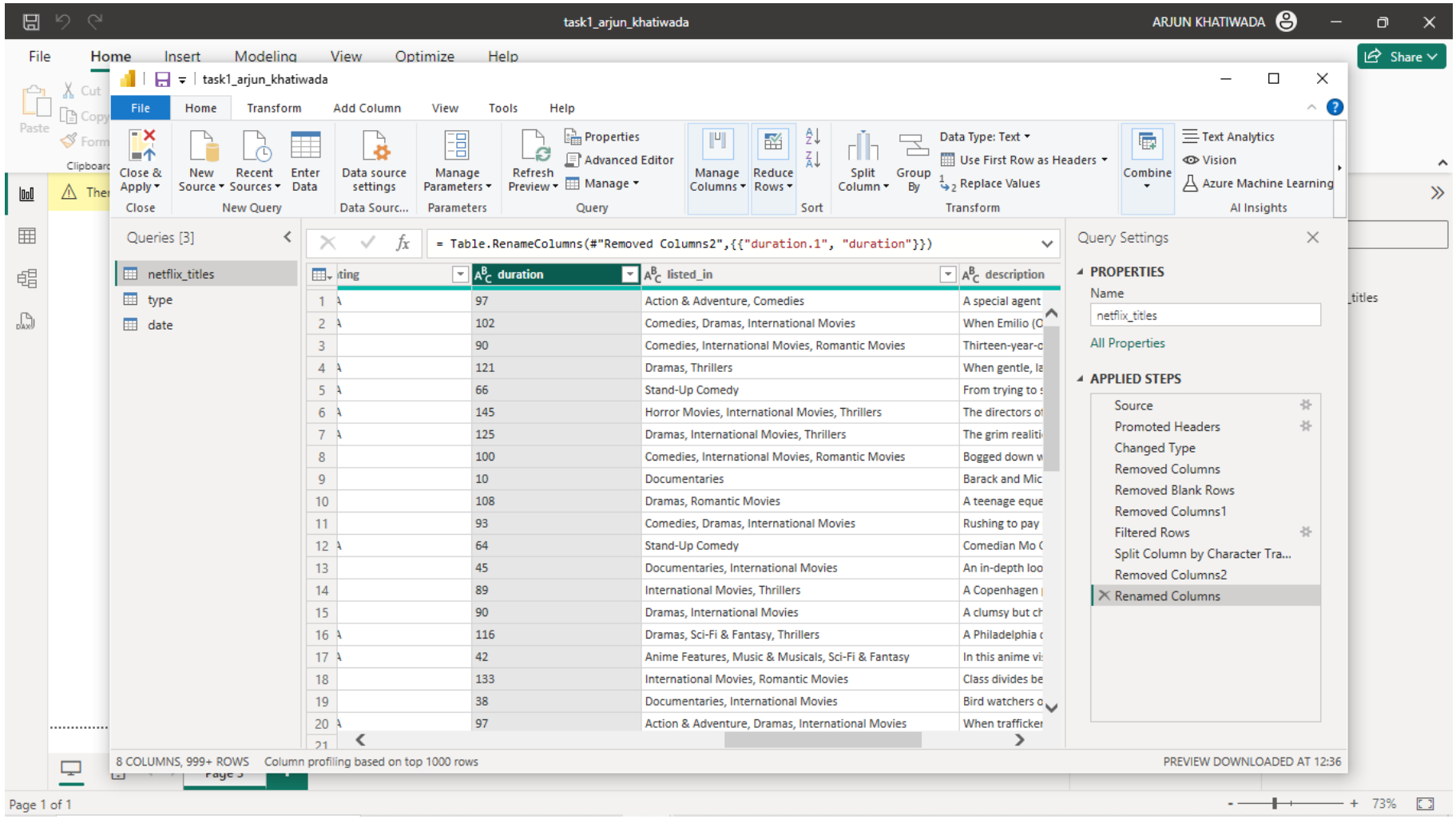
73%



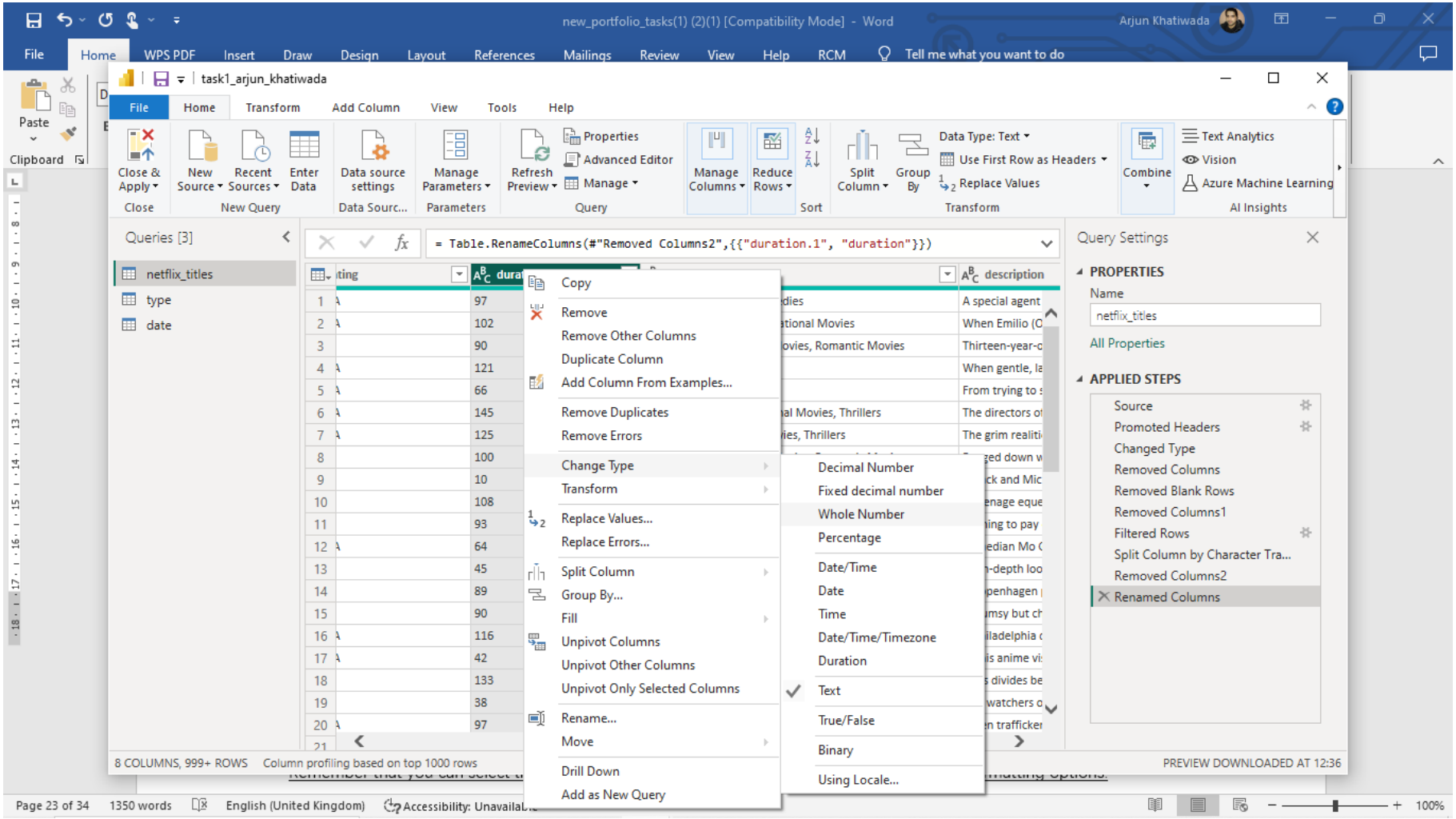
12) Removing the suffix ‘min’ from the column ‘duration’.

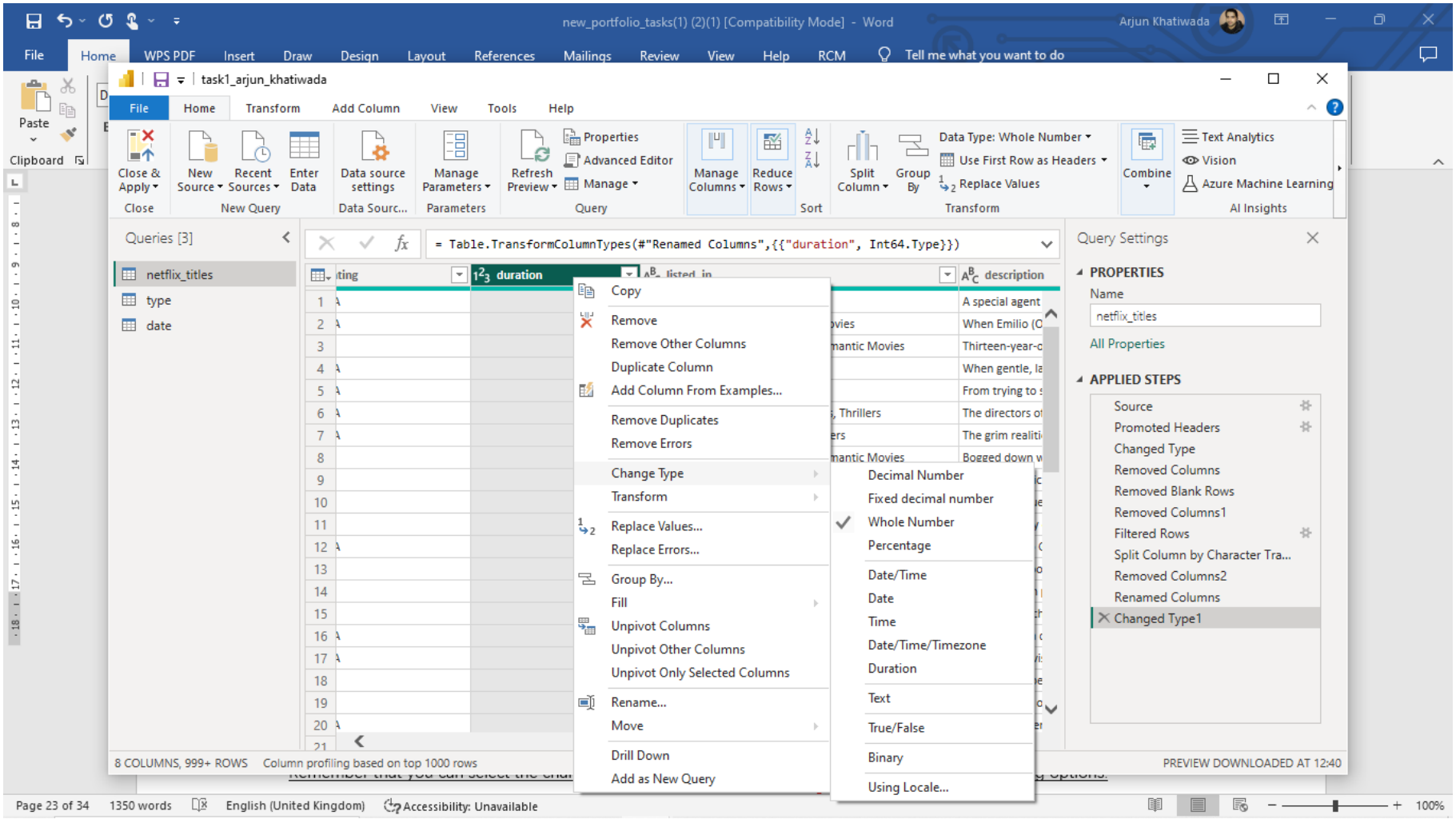

Screenshots
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Manage Columns
Reduce Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Combine
Text Analytics
Vision
Azure Machine Learning
Queries [3]
netflix_titles
type
date
= Table.SelectRows(#"Removed Columns1", each ([duration] <> "1 Season" and
duration
description
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Split Column
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
By Delimiter...
By Number of Characters...
By Positions...
By Lowercase to Uppercase
By Uppercase to Lowercase
By Digit to Non-Digit
By Non-Digit to Digit
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Removed Columns
Removed Blank Rows
Removed Columns1
Filtered Rows
8 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 12:30
18 items

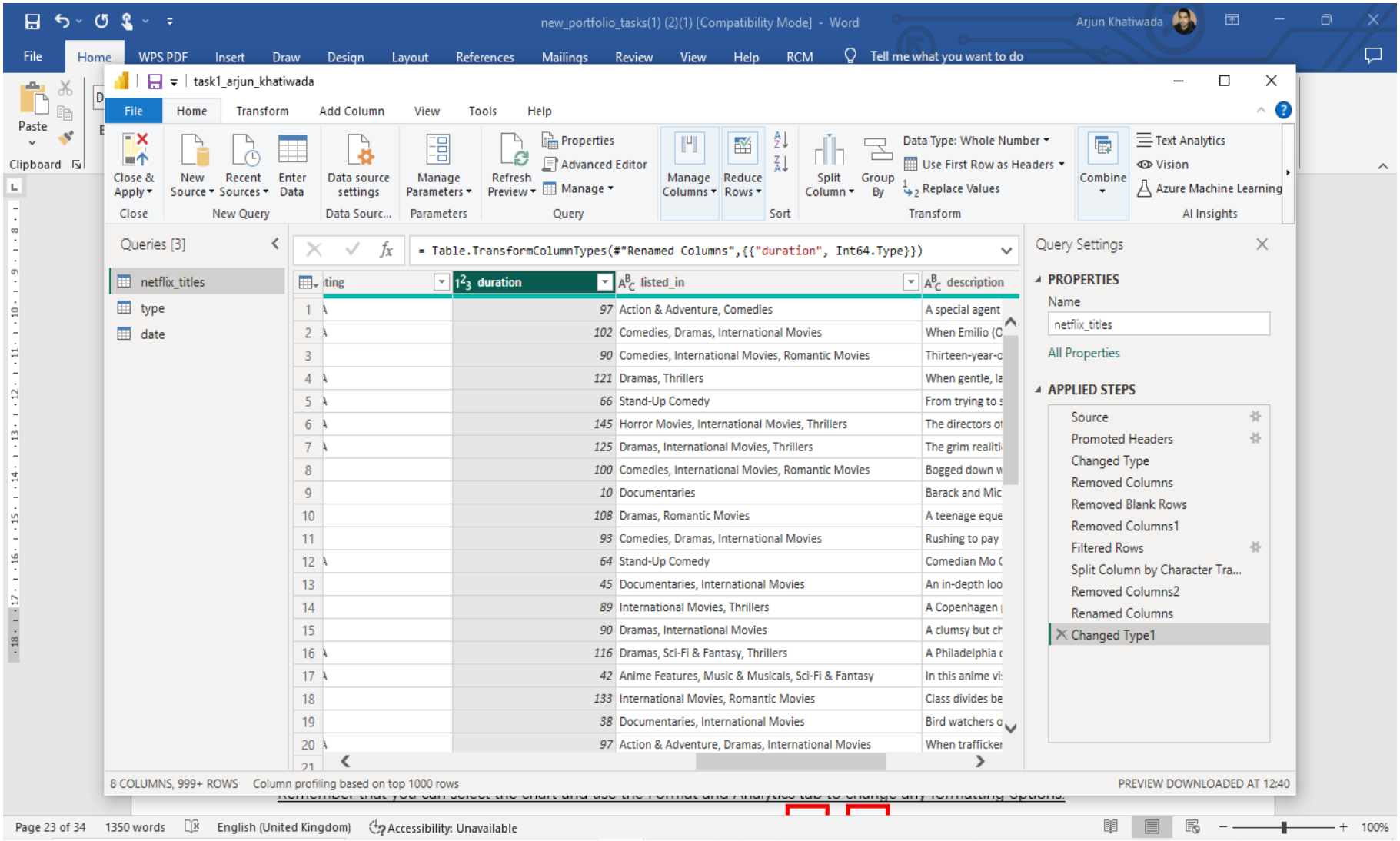

Manage
Screenshots
File
Home
Share
View
Picture Tools
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Manage Columns
Reduce Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Combine
Text Analytics
Vision
Azure Machine Learning
Close
New Query
Data Sourc...
Parameters
Query
Sort
Transform
AI Insights
Queries [3]
netflix_titles
type
date
= Table.SplitColumn(#"Filtered Rows", "duration",
duration.1
duration.2
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Split Column
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Removed Columns
Removed Blank Rows
Removed Columns1
Filtered Rows
Split Column by Character Tra...
9 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 12:35
20 items



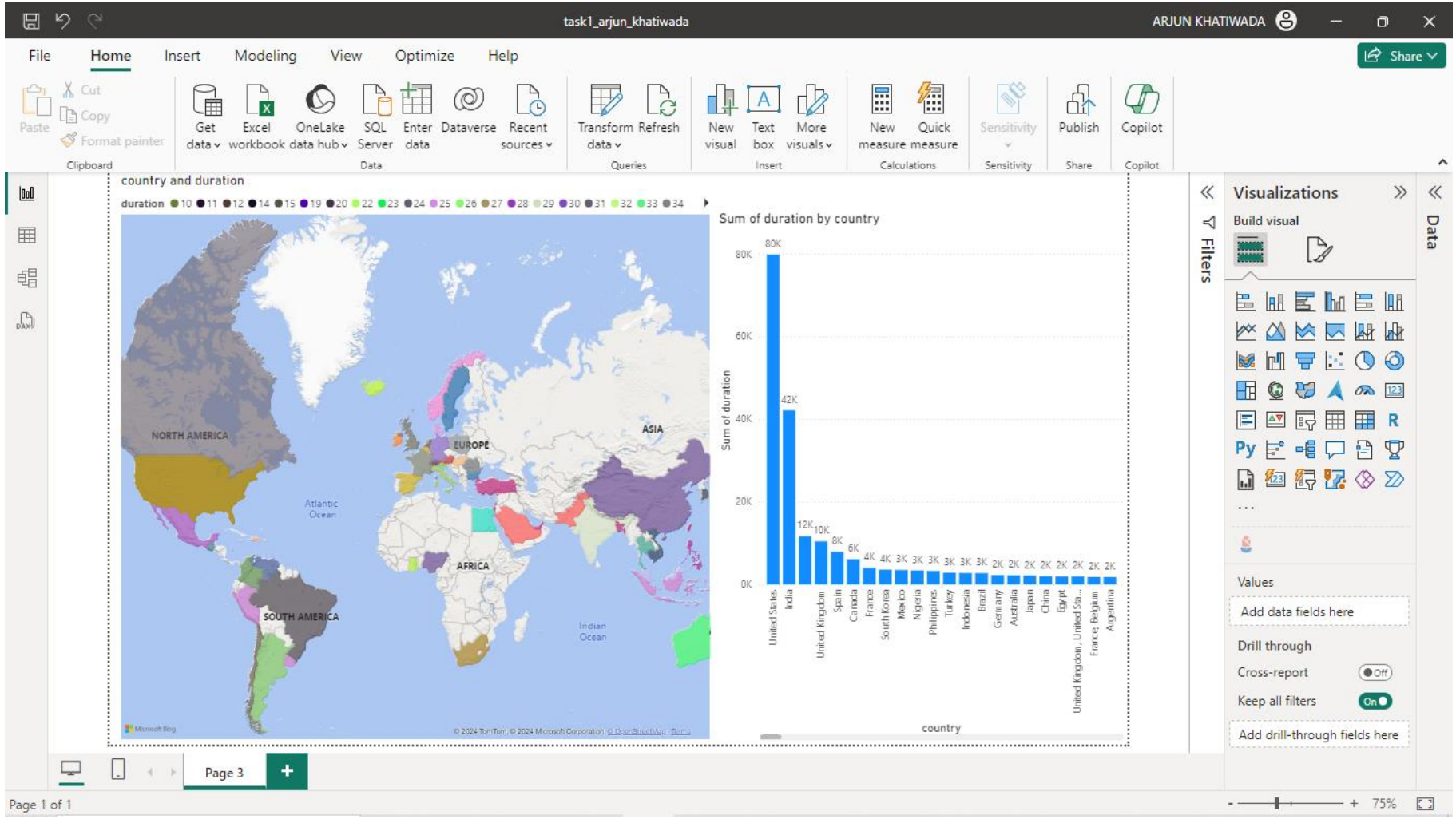
13) Changing datatype of column ‘duration’ to whole number


new_portfolio_tasks(1) (2)(1) [Compatibility Mode] - Word
Arjun Khatiwada
File
Home
WPS PDF
Insert
Draw
Design
Layout
References
Mailings
Review
View
Help
RCM
Tell me what you want to do
Paste
Clipboard
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
Close
New Source
Recent Sources
Enter Data
New Query
Data source settings
Data Sourc...
Manage Parameters
Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Query
Manage Columns
Reduce Rows
Sort
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Transform
Combine
Text Analytics
Vision
Azure Machine Learning
AI Insights
Queries [3]
netflix_titles
type
date
= Table.RenameColumns(#"Removed Columns2",{{"duration.1", "duration"}})
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Split Column
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
Decimal Number
Fixed decimal number
Whole Number
Percentage
Date/Time
Date
Time
Date/Time/Timezone
Duration
Text
True/False
Binary
Using Locale...
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Removed Columns
Removed Blank Rows
Removed Columns1
Filtered Rows
Split Column by Character Tra...
Removed Columns2
Renamed Columns
8 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 12:36
Page 23 of 34
1350 words
English (United Kingdom)
100%

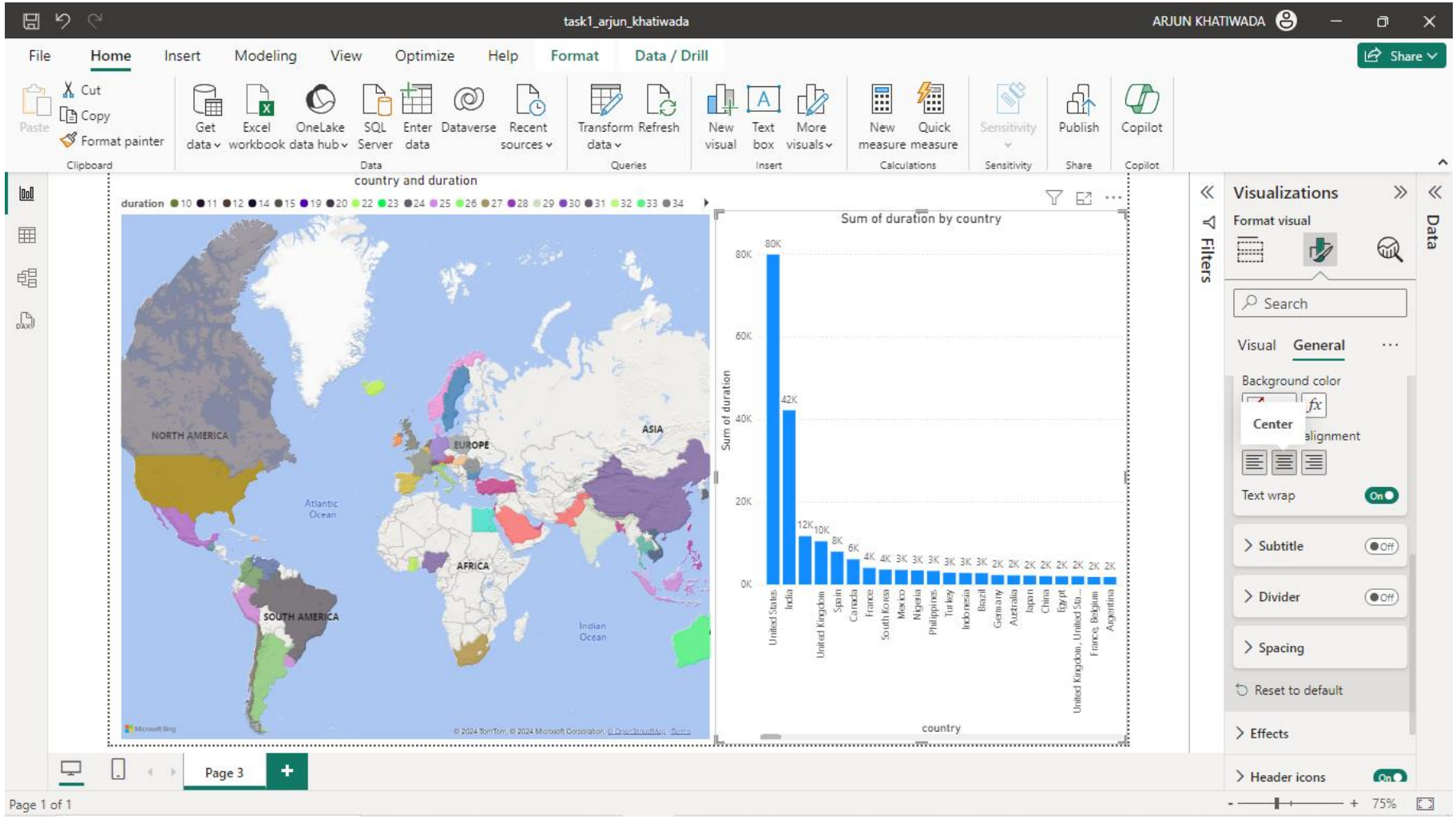

new_portfolio_tasks(1) (2)(1) [Compatibility Mode] - Word
Arjun Khatiwada
File
Home
WPS PDF
Insert
Draw
Design
Layout
References
Mailings
Review
View
Help
RCM
Tell me what you want to do
Paste
Clipboard
task1_arjun_khatiwada
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Manage Columns
Reduce Rows
Split Column
Group By
Data Type: Whole Number
Use First Row as Headers
Replace Values
Combine
Text Analytics
Vision
Azure Machine Learning
Close
New Query
Data Sourc...
Parameters
Query
Sort
Transform
AI Insights
Queries [3]
netflix_titles
type
date
= Table.TransformColumnTypes(#"Renamed Columns",{{"duration", Int64.Type}})
duration
description
Copy
Remove
Remove Other Columns
Duplicate Column
Add Column From Examples...
Remove Duplicates
Remove Errors
Change Type
Transform
Replace Values...
Replace Errors...
Group By...
Fill
Unpivot Columns
Unpivot Other Columns
Unpivot Only Selected Columns
Rename...
Move
Drill Down
Add as New Query
Decimal Number
Fixed decimal number
Whole Number
Percentage
Date/Time
Date
Time
Date/Time/Timezone
Duration
Text
True/False
Binary
Using Locale...
Query Settings
PROPERTIES
Name
netflix_titles
All Properties
APPLIED STEPS
Source
Promoted Headers
Changed Type
Removed Columns
Removed Blank Rows
Removed Columns1
Filtered Rows
Split Column by Character Tra...
Removed Columns2
Renamed Columns
Changed Type1
8 COLUMNS, 999+ ROWS
Column profiling based on top 1000 rows
PREVIEW DOWNLOADED AT 12:40
Page 23 of 34
1350 words
English (United Kingdom)
Accessibility: Unavailable
100%



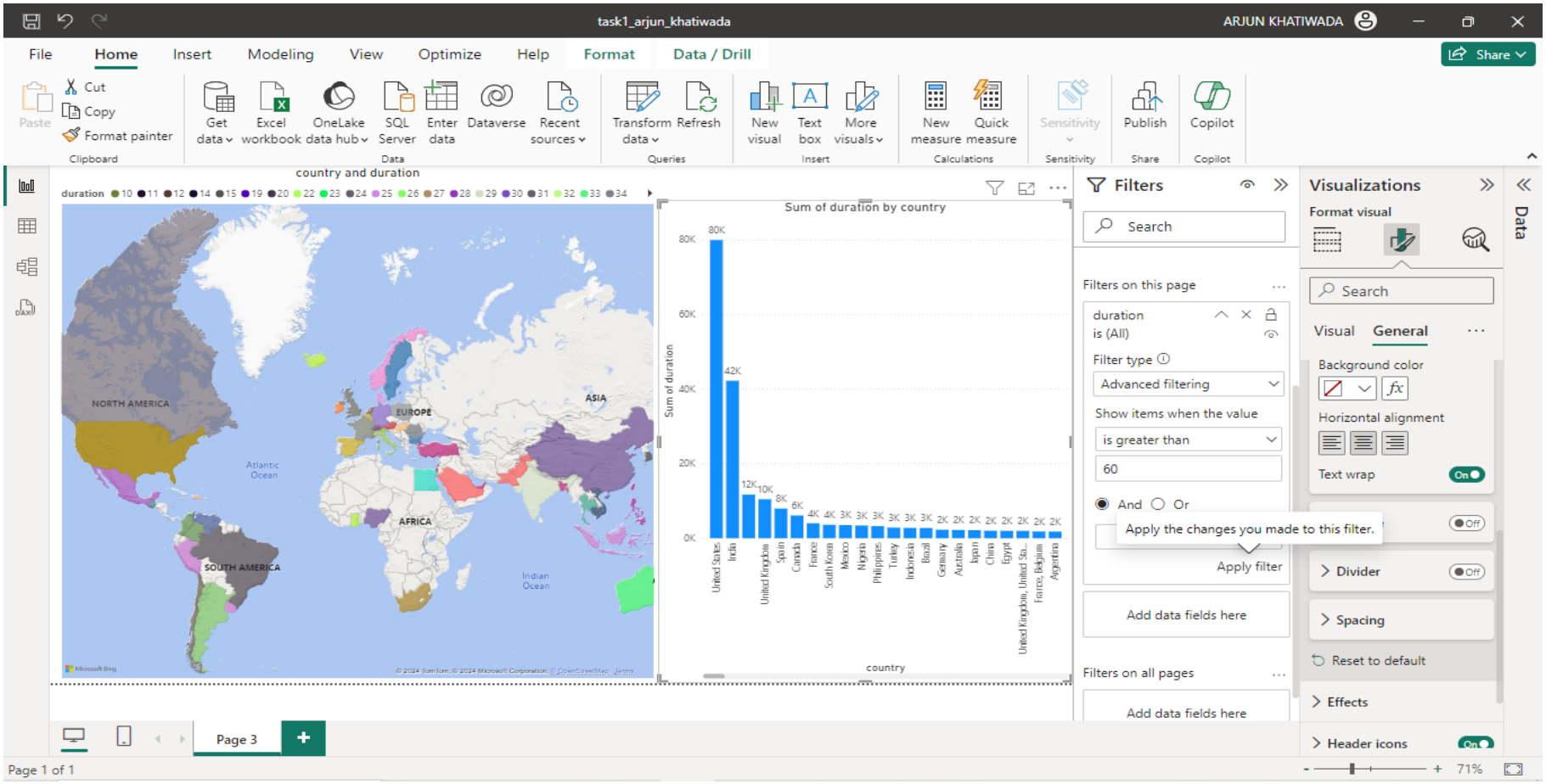
14) Map showing the duration by country

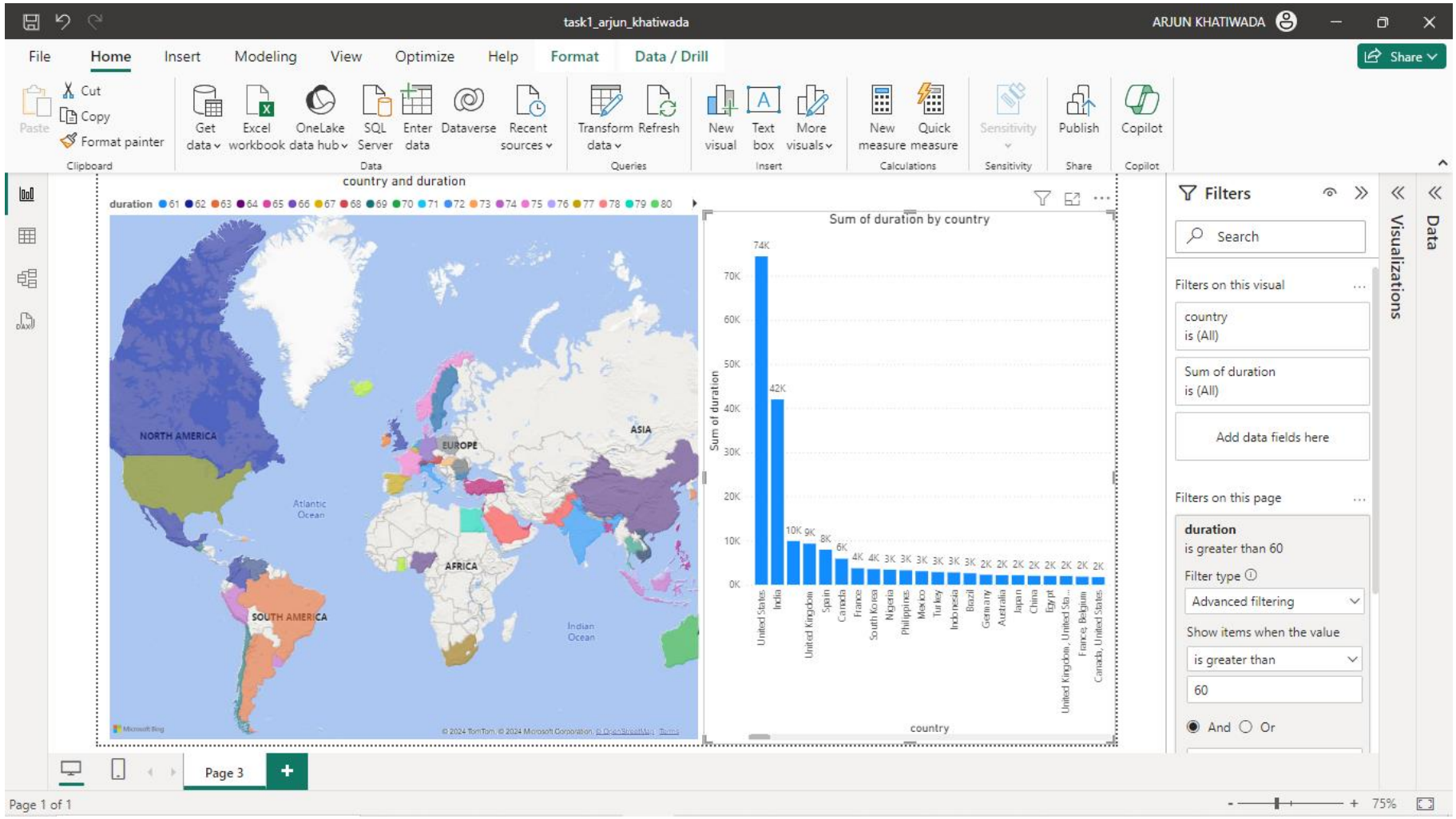


15) Centring the title

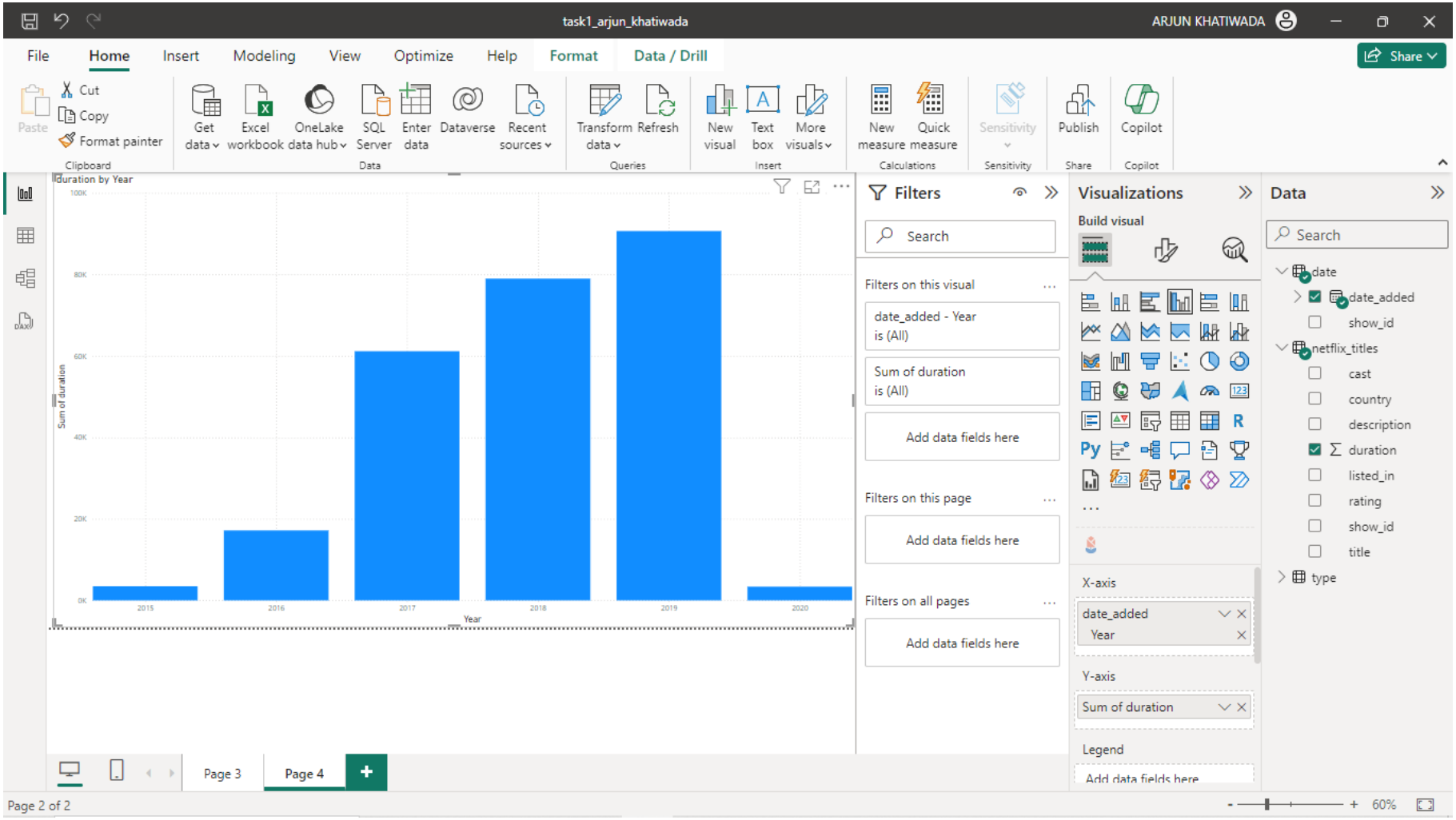


16) Filtering all the values to get ‘duration’ > 60 mins.

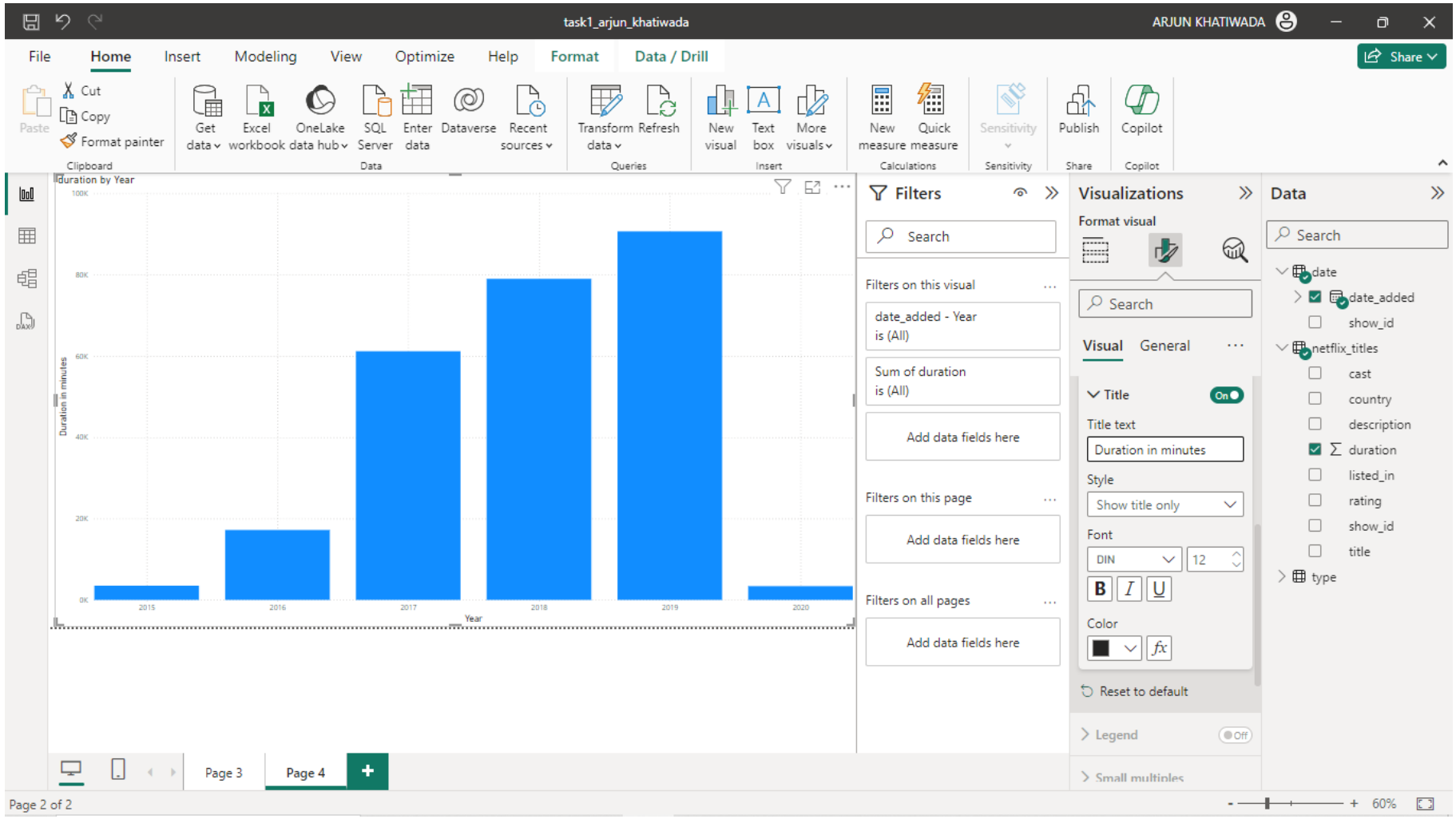

task1_arjun_khatiwada
ARJUN KHATIWADA
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Share
Cut
Copy
Format painter
Paste
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Publish
Share
Copilot
Copilot
country and duration
duration 10 11 12 14 15 19 20 22 23 24 25 26 27 28 29 30 31 32 33 34
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
Atlantic Ocean
Indian Ocean
Sum of duration by country
Sum of duration
80K
60K
40K
20K
0K
80K
42K
12K
10K
8K
6K
4K
4K
3K
3K
3K
3K
3K
3K
2K
2K
2K
2K
2K
2K
2K
2K
United States
India
United Kingdom
Spain
Canada
France
South Korea
Mexico
Nigeria
Philippines
Turkey
Indonesia
Brazil
Germany
Australia
Japan
China
Egypt
United Kingdom, United Sta...
France, Belgium
Argentina
country
Filters
Search
Filters on this page
duration
is (All)
Filter type
Advanced filtering
Show items when the value
is greater than
60
And
Or
Apply the changes you made to this filter.
Apply filter
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Format visual
Search
Visual
General
Background color
Horizontal alignment
Text wrap
On
Off
Divider
Off
Spacing
Reset to default
Effects
Header icons
On
Data
Page 3
Page 1 of 1
71%



17) Chart showing 'duration by year'.



18) Changing y-axis title to 'Duration in minutes'.

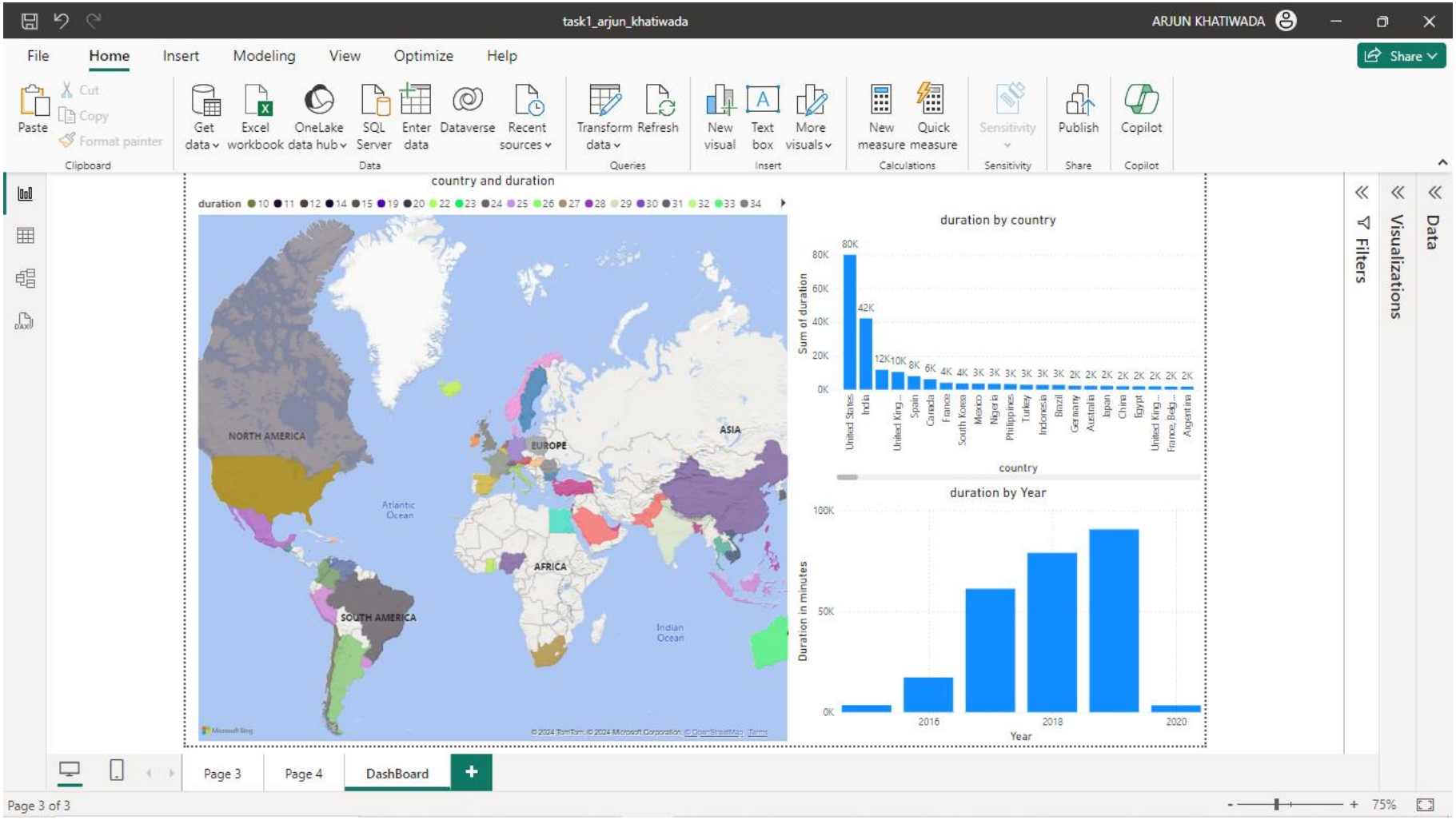


Dashboard and Model View:

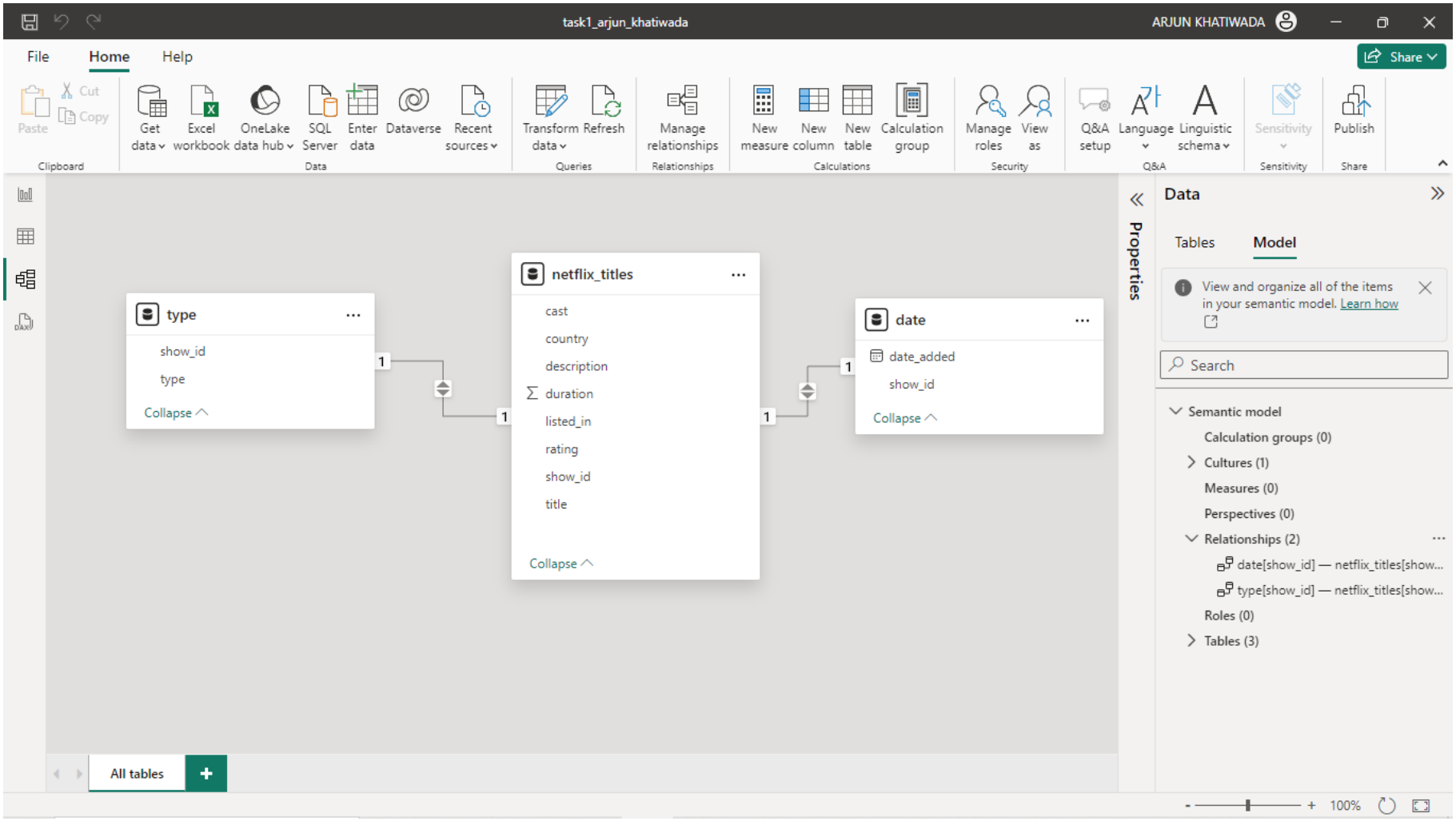
Link to Dashboard: [big_data_portfolio_task1_arjun_khatiwada]

Screenshots:

Dashboard:

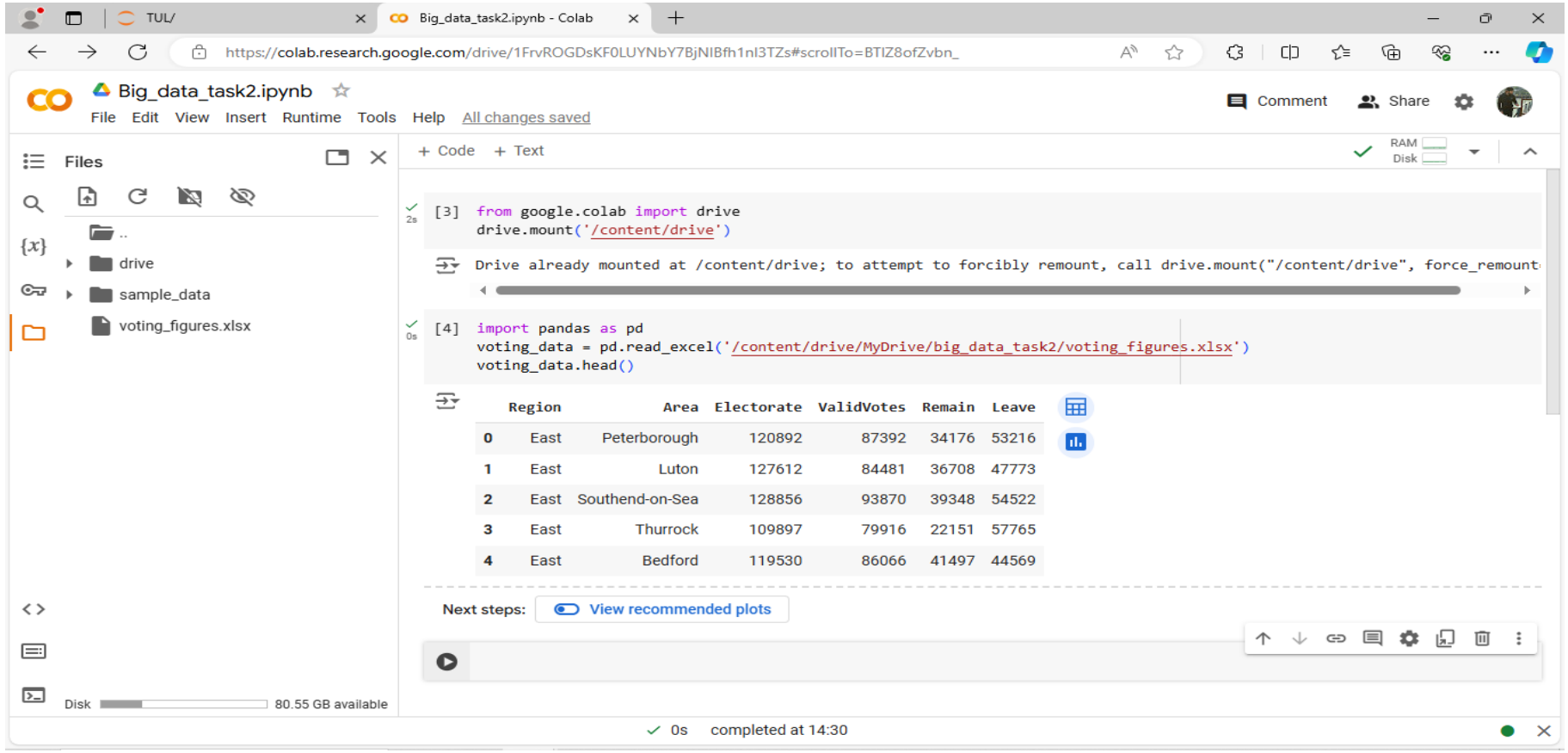


Model View:

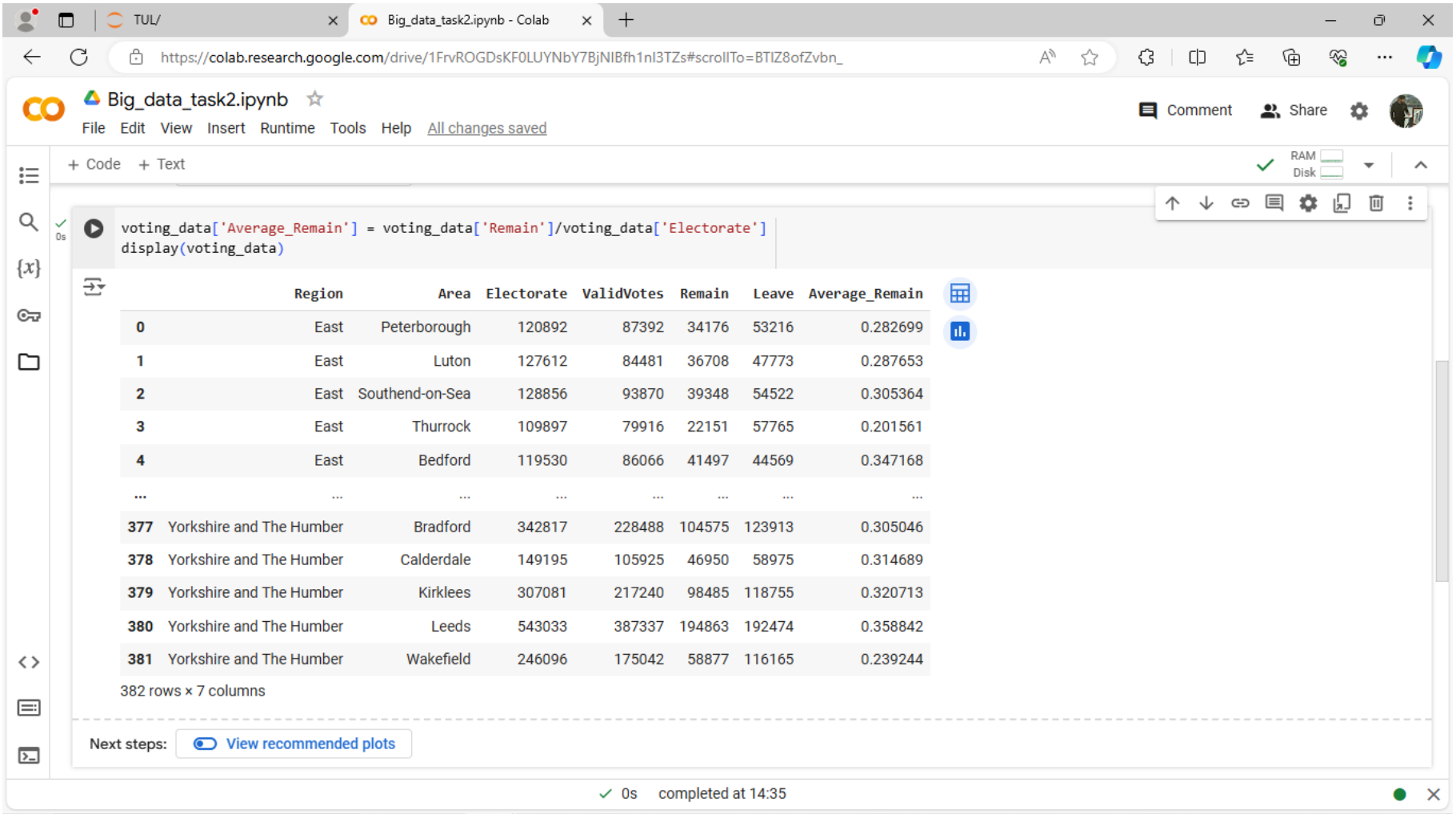

task1_arjun_khatiwada
ARJUN KHATIWADA
File
Home
Help
Share
Paste
Cut
Copy
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
Manage relationships
Relationships
New measure
New column
New table
Calculation group
Calculations
Manage roles
View as
Security
Q&A setup
Language
Linguistic schema
Q&A
Sensitivity
Publish
Share
type
show_id
type
Collapse
netflix_titles
cast
country
description
duration
listed_in
rating
show_id
title
Collapse
date
date_added
show_id
Collapse
1
1
1
1
Properties
Data
Tables
Model
View and organize all of the items in your semantic model. Learn how
Search
Semantic model
Calculation groups (0)
Cultures (1)
Measures (0)
Perspectives (0)
Relationships (2)
date[show_id] — netflix_titles[show...
type[show_id] — netflix_titles[show...
Roles (0)
Tables (3)
All tables
100%

Task 2: Python and Pandas (20 points)

1) Importing the excel dataset voting figures into google collab using pandas



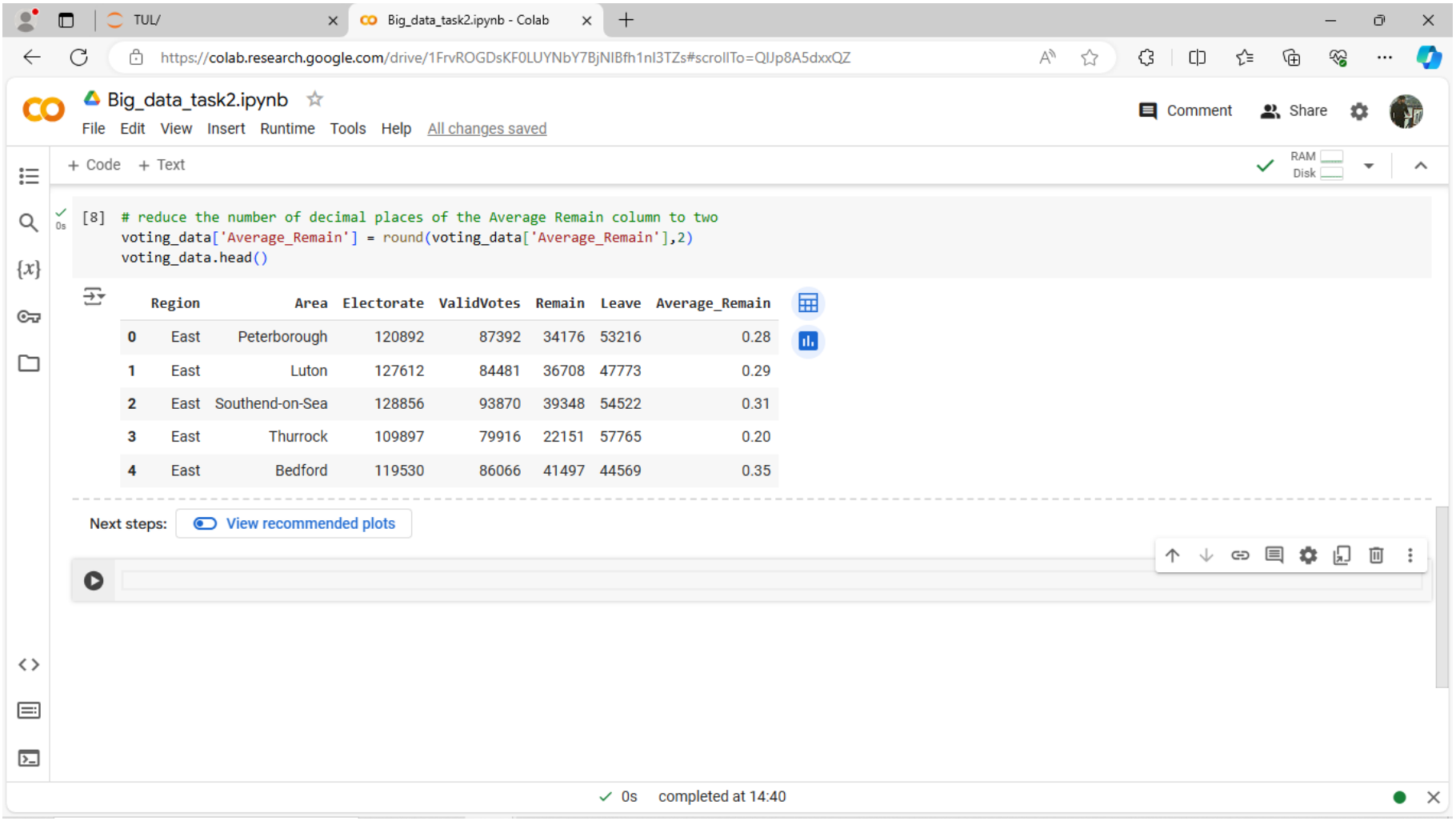
2) Adding new column 'Average_Remain' which is proportion of remain in total electorate



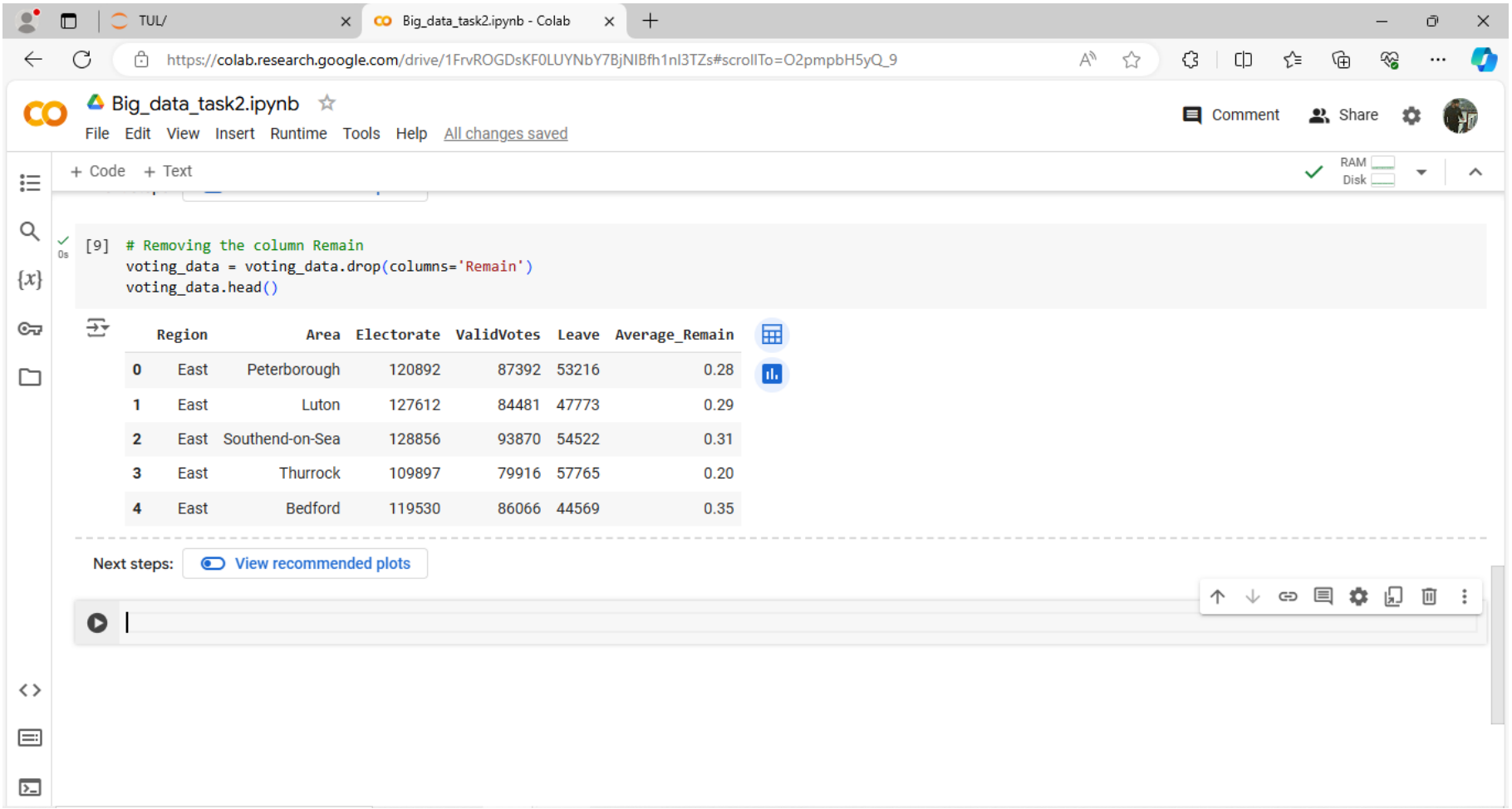
	Region	Area	Electorate	ValidVotes	Remain	Leave	Average_Remain
0	East	Peterborough	120892	87392	34176	53216	0.282699
1	East	Luton	127612	84481	36708	47773	0.287653
2	East	Southend-on-Sea	128856	93870	39348	54522	0.305364
3	East	Thurrock	109897	79916	22151	57765	0.201561
4	East	Bedford	119530	86066	41497	44569	0.347168
...	...	...	...	...	...	...	...
377	Yorkshire and The Humber	Bradford	342817	228488	104575	123913	0.305046
378	Yorkshire and The Humber	Calderdale	149195	105925	46950	58975	0.314689
379	Yorkshire and The Humber	Kirklees	307081	217240	98485	118755	0.320713
380	Yorkshire and The Humber	Leeds	543033	387337	194863	192474	0.358842
381	Yorkshire and The Humber	Wakefield	246096	175042	58877	116165	0.239244

382 rows × 7 columns

3) Rounding up newly added column to 2 decimal places



4) Removing the column Remain



	Region	Area	Electorate	ValidVotes	Leave	Average_Remain
0	East	Peterborough	120892	87392	53216	0.28
1	East	Luton	127612	84481	47773	0.29
2	East	Southend-on-Sea	128856	93870	54522	0.31
3	East	Thurrock	109897	79916	57765	0.20
4	East	Bedford	119530	86066	44569	0.35

5) Creating new column 'Remain_Above_Average' to classify whether the remain proportion is greater than average remain proportion or not.

```
# We want to create a column "Remain Above Average"
voting_data['Remain_Above_Average'] = ['Yes' if x >= voting_data['Average_Remain'].mean() else 'No' for x in voting_data['Average_Remain']]
voting_data.head()
```

	Region	Area	Electorate	ValidVotes	Leave	Average_Remain	Remain_Above_Average
0	East	Peterborough	120892	87392	53216	0.28	No
1	East	Luton	127612	84481	47773	0.29	No
2	East	Southend-on-Sea	128856	93870	54522	0.31	No
3	East	Thurrock	109897	79916	57765	0.20	No
4	East	Bedford	119530	86066	44569	0.35	Yes

Next steps: View recommended plots

0s completed at 14:43

6) Proportion of Remain Region wise

TUL/ | Big_data_task2.ipynb - Colab

https://colab.research.google.com/drive/1FrvROGDsKF0LUYNbY7BjNIBfh1nI3TZs#scrollTo=x1up4e_L0AwS

Big_data_task2.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

```
# Remain proportion across each Region of the UK
voting_data['Remain_Votes'] = voting_data['ValidVotes'] - voting_data['Leave'] #calculating remain votes
voting_data['Remain_votes_proportion'] = round((voting_data['ValidVotes'] - voting_data['Leave'])/voting_data['ValidVotes'],2)
voting_data_for_prop = voting_data[['Region','ValidVotes','Remain_Votes']]
voting_data_groupby = voting_data_for_prop.groupby('Region')[['Remain_Votes','ValidVotes']].sum()
voting_data_groupby['RemainVotes_Proportion'] = round(voting_data_groupby['Remain_Votes']/voting_data_groupby['ValidVotes'],2)
voting_data_groupby
```

Region	Remain_Votes	ValidVotes	RemainVotes_Proportion
East	1448616	3328983	0.44
East Midlands	1033036	2508515	0.41
London	2263519	3776751	0.60
North East	562595	1340698	0.42
North West	1699020	3665945	0.46
Northern Ireland	440707	790149	0.56
Scotland	1661191	2679513	0.62
South East	2391718	4959683	0.48
South West	1503019	3172730	0.47
Wales	772347	1626919	0.47
West Midlands	1207175	2962862	0.41
Yorkshire and The Humber	1158298	2739235	0.42

✓ 0s completed at 14:52

Questions		Provide the code for each question

1	Import the data into pandas	`from google.colab import drive`
		`drive.mount('/content/drive')`
		`import pandas as pd`
		`voting_data = pd.read_excel('/content/drive/MyDrive/big_data_task2/voting_figures.xlsx')`
		`voting_data.head()`

2	Add average remain column	`voting_data['Average_Remain'] = voting_data['Remain']/voting_data['Electorate']` `display(voting_data)`

3	Reduce to 2 decimal places.	`# reduce the number of decimal places of the Average Remain column to two`
		`voting_data['Average_Remain'] = round(voting_data['Average_Remain'],2)`
		`voting_data.head()`

4	Remove the Remain column.	`# Removing the column Remain`
		`voting_data = voting_data.drop(columns='Remain')`
		`voting_data.head()`

5	Create the transformed column Remain Above Average.	`# We want to create a column "Remain Above Average"`
		`voting_data['Remain_Above_Average'] = ['Yes' if x >= voting_data['Average_Remain'].mean() else 'No' for x in voting_data['Average_Remain']]`
		`voting_data.head()`

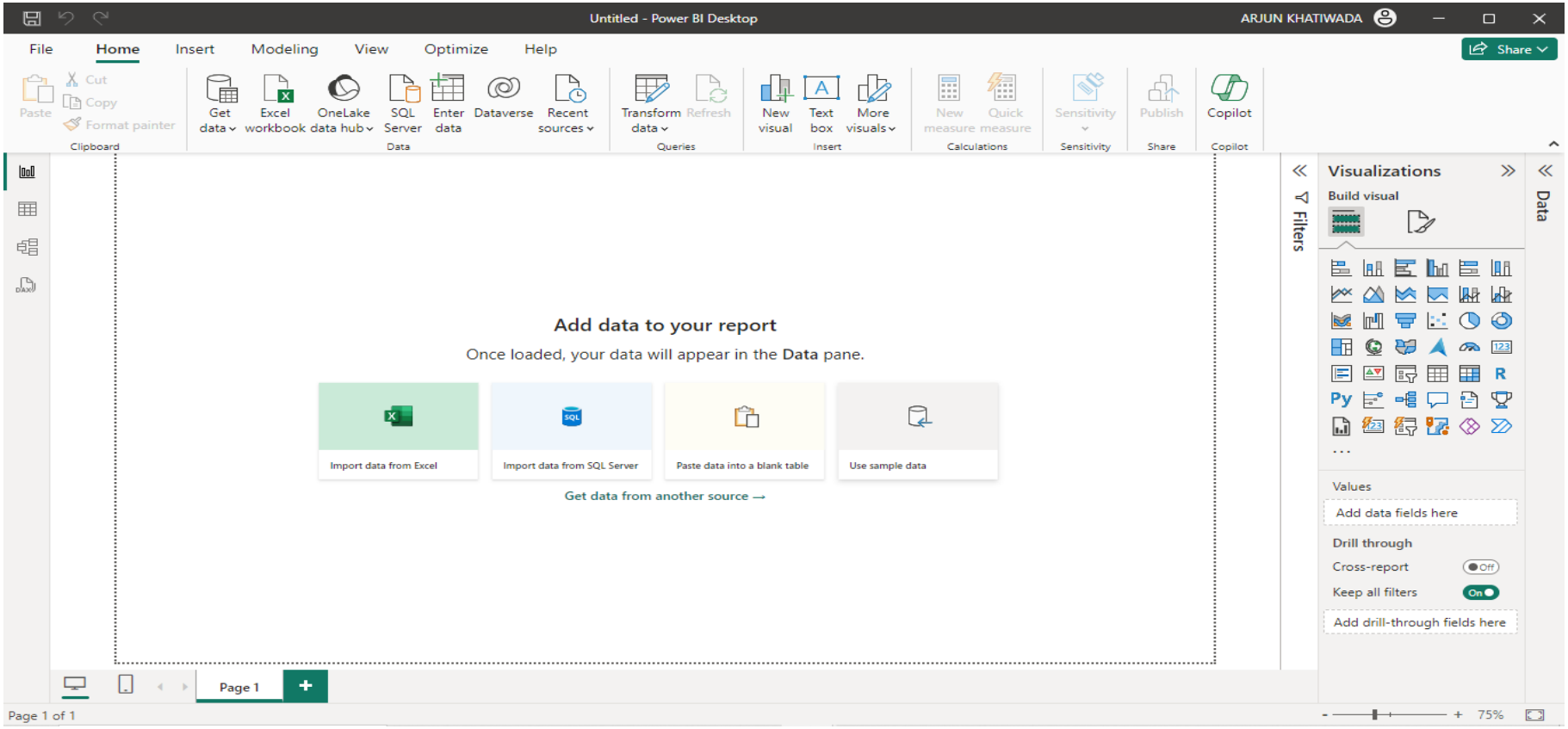
6	Calculate the Remain proportion for each region of the UK.	`# Remain proportion across each Region of the UK` `voting_data['Remain_Votes'] = voting_data['ValidVotes'] - voting_data['Leave'] #calculating remain votes`

		`voting_data['Remain_votes_proportion'] = round((voting_data['ValidVotes'] - voting_data['Leave'])/voting_data['ValidVotes'],2)`
		`voting_data_for_prop = voting_data[['Region','ValidVotes','Remain_Votes']]`
		`voting_data_groupby = voting_data_for_prop.groupby('Region')[['Remain_Votes','ValidVotes']].sum()`
		`voting_data_groupby['RemainVotes_Proportion'] = round(voting_data_groupby['Remain_Votes']/voting_data_groupby['ValidVotes'],2)`
		`voting_data_groupby`

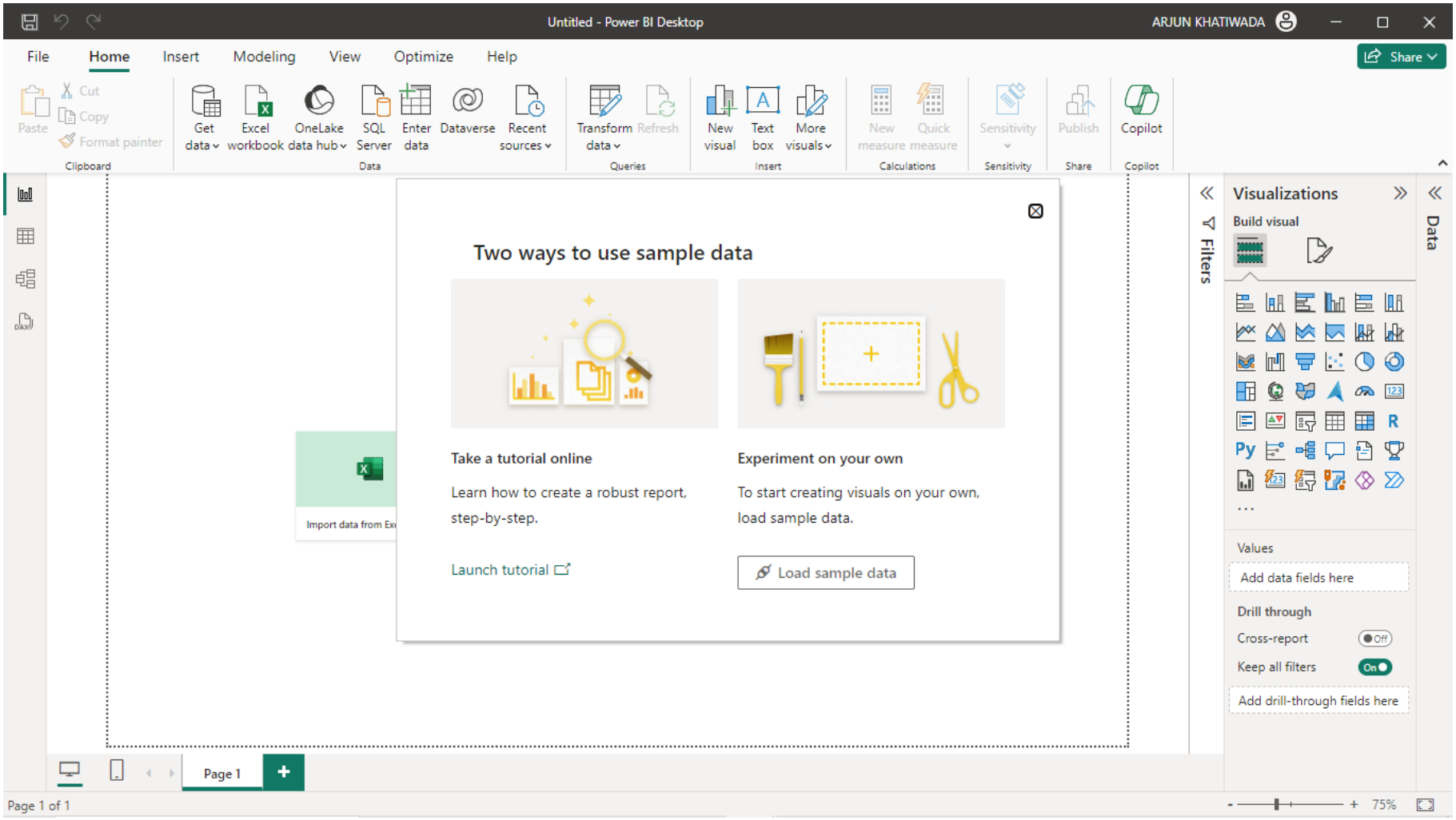
Link to google Collab notebook: [big_data_task2_arjun_khatiwada_collab_notebook]

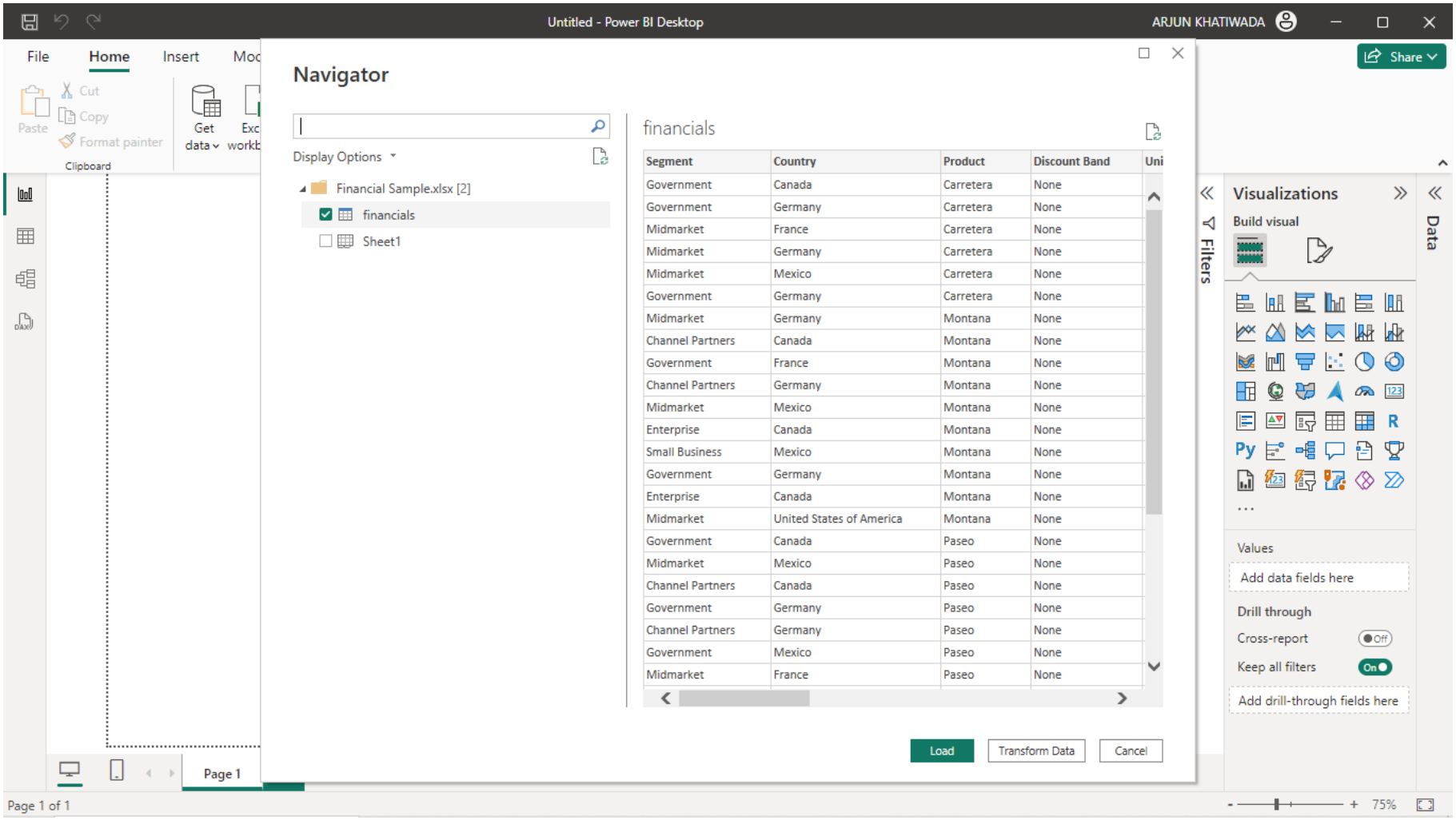
Task 3: Dashboard Design (20 points total)

Using a financial dataset from sample dataset in power BI.



Loading the sample data.

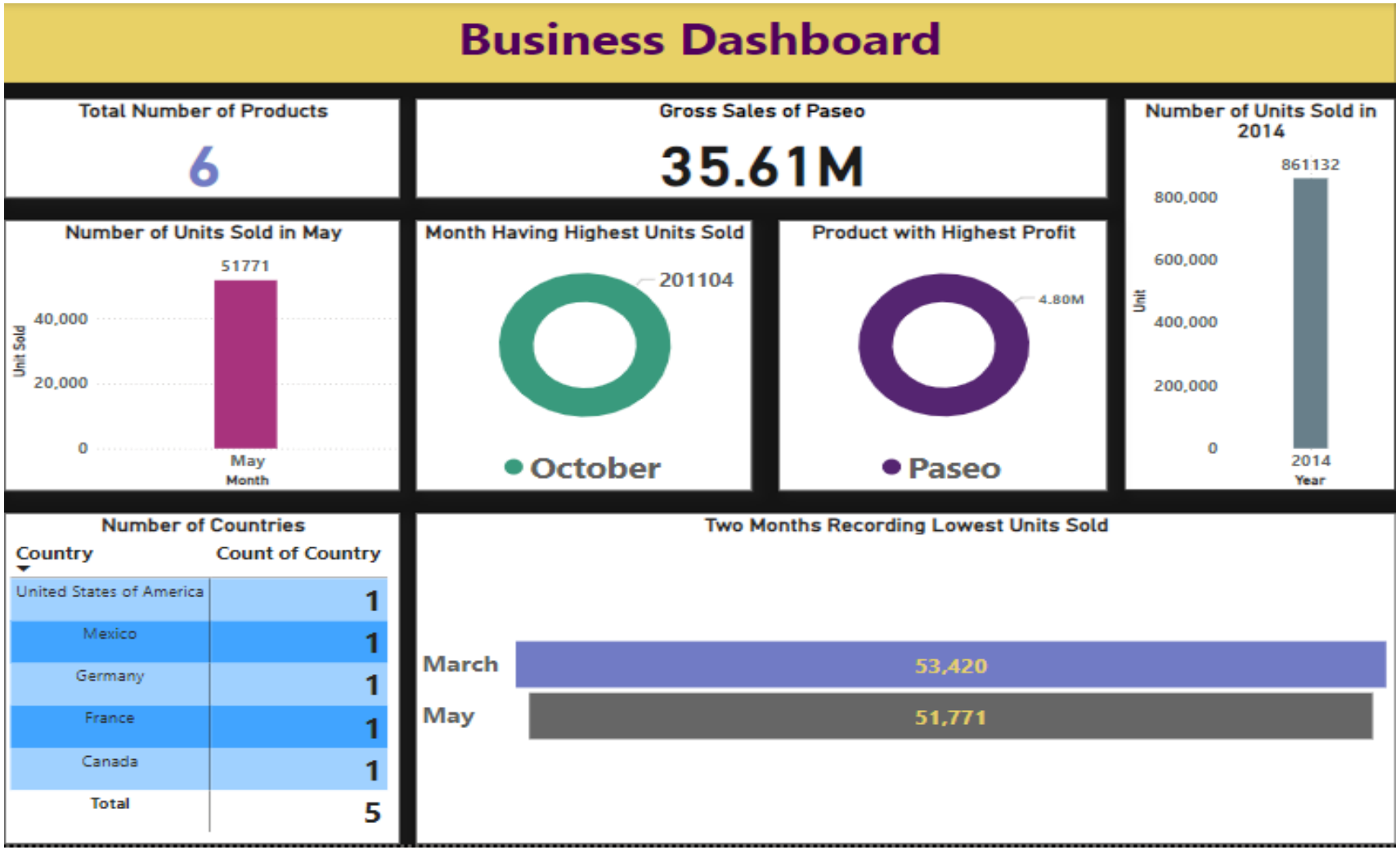

Untitled - Power BI Desktop
ARJUN KHATIWADA
File
Home
Insert
Modeling
View
Optimize
Help
Share
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
OneLake data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Publish
Share
Copilot
Copilot
Two ways to use sample data
Take a tutorial online
Learn how to create a robust report, step-by-step.
Launch tutorial
Experiment on your own
To start creating visuals on your own, load sample data.
Load sample data
Import data from Ex
Filters
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Data
Page 1
Page 1 of 1
75%


Untitled - Power BI Desktop
ARJUN KHATIWADA
File
Home
Insert
Share
Cut
Copy
Paste
Format painter
Clipboard
Get data
Navigator
Display Options
Financial Sample.xlsx [2]
financials
Sheet1
financials
Segment | Country | Product | Discount Band
Government | Canada | Carretera | None
Government | Germany | Carretera | None
Midmarket | France | Carretera | None
Midmarket | Germany | Carretera | None
Midmarket | Mexico | Carretera | None
Government | Germany | Carretera | None
Midmarket | Germany | Montana | None
Channel Partners | Canada | Montana | None
Government | France | Montana | None
Channel Partners | Germany | Montana | None
Midmarket | Mexico | Montana | None
Enterprise | Canada | Montana | None
Small Business | Mexico | Montana | None
Government | Germany | Montana | None
Enterprise | Canada | Montana | None
Midmarket | United States of America | Montana | None
Government | Canada | Paseo | None
Midmarket | Mexico | Paseo | None
Channel Partners | Canada | Paseo | None
Government | Germany | Paseo | None
Channel Partners | Germany | Paseo | None
Government | Mexico | Paseo | None
Midmarket | France | Paseo | None
Load
Transform Data
Cancel
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here
Filters
Data
Page 1
Page 1 of 1
75%

	Business Questions	Answer
1	Which product had the highest profit?	**Paseo**
2	How many different products are in the dataset?	**6**
3	What was the gross sale associated with the "Paseo" product?	**35.61 million**
4	How many units were sold in 2014?	**861132**
5	How many units were sold in May (over the two years)?	**51771**
6	Which month had the highest number of units sold?	**October**
7	How many different countries are in the dataset?	**5**
8	Which two months had the lowest number of units sold?	**May and March**
	Arrange your tiles/charts to create a well-designed dashboard (use the knowledge covered in the lessons). Add a title, change the colours/formatting options of the charts. Once you are happy with the dashboard design, **paste either a screenshot or a link to a screenshot e.g. OneDrive.**	

Link to the Dashboard: business_dashboard_task3_arjun_khatiwada

Screenshot:



Theoretical aspects of visualization

1.Various data sets are denoted by different colours; for example, teal indicates the month with the highest number of units sold, and purple indicates the number of units sold in May. This fits perfectly with the Gestalt principle of similarity, which groups related objects together based on similar colours to facilitate rapid visual appeal and discrimination.

2. For example, the dashboard uses bar charts to show the "Number of Units Sold in May" and donut charts to show the "Month Having Highest Units Sold." Applying the Gestalt principle of figure-ground where the chart style makes the data stand out as the figure against the background ensures clarity and efficacy in information conveyance.

3. Important statistics like "Total Number of Products" and "Gross Sales of Paseo" are displayed clearly and prominently. According to the Gestalt concept of simplicity, which emphasises clear and simple presentations, simplifying data representation lowers cognitive load and enables users to quickly understand important information.

Task 4: Data Misrepresentation (25 points)

In my approach, as a data analyst confronted with the ethical dilemma of data falsification, I would prioritise professionalism, accountability, honesty, and societal impact.

Responsibility and Professionalism
The first thing I would do would be to carefully examine the information to make sure I understood everything. Preventing unintentional mistakes that could exacerbate the moral conundrum is crucial. To establish an open dialogue with my supervisors, I must thoroughly record the advantages and disadvantages of data manipulation. Beyond the obvious financial ramifications, my duty also includes preserving the company's long-term interests and standing in the marketplace (National Association of Social Workers,).

Integrity
The foundation of moral data practices is integrity. I would bring my research to the attention of those in charge, stressing the value of being truthful and upholding confidence. Emphasising the long-term advantages of moral behavior like steady growth, devoted clients, and a positive brand image—aligns with moral principles and fosters an integrity-based culture within the company. This method emphasises that morality and strategic advantage go hand in hand with ethical behaviour (National Association of Social Workers,).

Social Impact
It is crucial to take the social repercussions of data manipulation into account. I am aware that unfavourable societal effects could damage

the business's standing and undermine public confidence, particularly in a market with as much scrutiny as the UK. In talks with management, I would draw attention to these effects and push for moral behaviour that takes into account wider societal ramifications, such as following UK-specific laws and public expectations. This campaigning serves to emphasise how important social responsibility is (National Association of Social Workers,).

In conclusion, a strong commitment to professionalism, accountability, integrity, and social impact is necessary for data analysts to successfully navigate the problem of data misrepresentation. I uphold the highest standards of my profession by carefully examining the facts, recording risks and benefits, supporting moral behaviour, and raising issues as needed. This strategy upholds a socially conscious and ethical culture while safeguarding the long-term interests of my business.

Task 5: Statistics (15 points)

5.1) Certain metrics of central tendency and dispersion are more suited when working with a dataset that contains major outliers since they are less impacted by extreme values. The suggested actions are as follows:

Median (Central Tendency): When a dataset is sorted from lowest to highest value, the midway value is known as the median. The median is a more reliable indicator of central tendency in a set of large outliers than the mean since it is unaffected by outliers (Laaser *et al.*, 2013).

Interquartile Range (IQR) (Dispersion): The spread of the middle 50% of the data is measured by the IQR. The difference between the first quartile (Q1) and the third quartile (Q3) is used to compute it. The IQR is a better way to quantify dispersion in datasets with outliers since it concentrates on the centre region of the data and is unaffected by extreme values (Laaser *et al.*, 2013).

Median Absolute Deviation (MAD) (Dispersion): The median of the absolute deviations from the dataset's median is represented by the MAD, another reliable measure of variability. This measure works well with datasets that contain significant outliers since it is less affected by outliers than standard deviation (Laaser *et al.*, 2013).

Example

Let's consider a dataset with significant outliers to illustrate the appropriate measures of central tendency and dispersion:

+ Code + Text

```python
import numpy as np
from scipy import stats

# Example Dataset
data = np.array([10, 12, 14, 15, 15, 16, 18, 18, 20, 22, 150])

# Mean
mean = np.mean(data)

# Median
median = np.median(data)

# Standard Deviation
std_dev = np.std(data, ddof=1)

# Interquartile Range (IQR)
q75, q25 = np.percentile(data, [75 ,25])
iqr = q75 - q25

# Median Absolute Deviation (MAD)
mad = stats.median_abs_deviation(data)

# Displaying the results
print(f"Mean: {mean:.2f} (affected by outlier)")
print(f"Median: {median} (not affected by outlier)")
print(f"Standard Deviation: {std_dev:.2f} (affected by outlier)")
print(f"IQR: {iqr} (not affected by outliers)")
print(f"MAD: {mad} (not affected by outliers)")
```

```
Mean: 28.18 (affected by outlier)
Median: 16.0 (not affected by outlier)
Standard Deviation: 40.55 (affected by outlier)
IQR: 4.5 (not affected by outliers)
MAD: 2.0 (not affected by outliers)
```

Conclusion:

Central Tendency: Median (16) and Dispersion: IQR (4.5) and MAD (2)

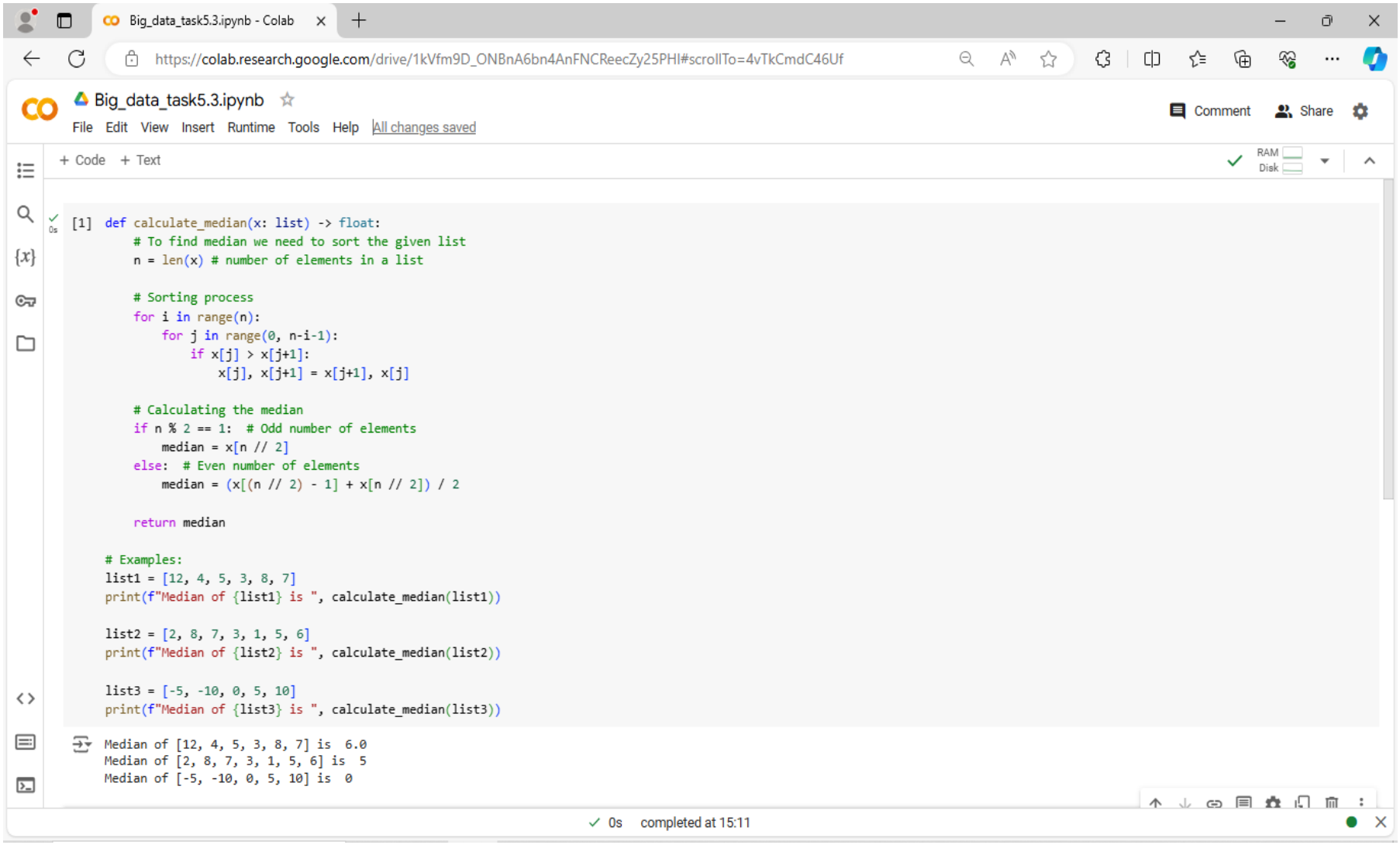
These measures (median, IQR, and MAD) are more appropriate than the mean and standard deviation when dealing with datasets containing significant outliers, as they provide a more accurate reflection of the central tendency and dispersion of the data.

5.2) There are various methodological problems with using a survey that depends on contacting a registry of landlines early on Monday morning to draw conclusions about the UK voting public. The following five main problems and possible fixes are listed:

Issue	Problem	Improvement	Relevant Examples
Coverage Bias	Samples have been impacted by the decline in landline ownership, particularly among younger and urban groups.	Use a blend of online, mobile, and landline surveys to guarantee a more representative sample.	2017 UK General Election: Because landline surveys were used in several of the polls, young turnout was underreported (BBC News, 2018).
Timing Bias	The survey being conducted on a Monday morning may exclude people with various schedules, students, and those who are employed.	Taking the survey during different times of day and also on other days of the week.	UK General Election: Because of scheduling bias that affected working voters, "Shy Tory Factor" polls understated Conservative support (Ashfold, 2014).
Nonresponse Bias	Results could be skewed since respondents who were available on a Monday morning might have had different traits and viewpoints.	To boost response rates, make follow-up attempts at various times and over a variety of channels (phone, internet).	Brexit Referendum 2016: Due to low response rates at different points in time, certain demographic groups were underrepresented (Rogers, 2016).
Questionnaire Design and Administration	Bias can be introduced through pollster interactions and question wording.	Make sure the questions are framed impartially, and instruct pollsters on normal operating procedures. Apply pretesting.	2010 UK General Election: A number of surveys drew criticism for using provocative questions that might elicit biassed answers (Pickup *et al.*, 2011).
Sampling Frame Limitations	Relying solely on landlines disqualifies recent arrivals, individuals without landlines, and those who would rather use mobile phones.	Include internet registrations and mobile phone directories in the expanded sampling frame. RDD, or random digit dialling, is used.	2015 UK General Election: Outdated sample frames that excluded families with only mobile phones contributed to some of the polling mistakes (Sturgis *et al.*, 2017).

The poll can produce more accurate and representative results by taking into account past instances from the UK and correcting these problems with the recommended adjustments.

5.3) Python function to calculate Median of any given array



Link to the google collab notebook: [big_data_task5.3_arjun_khatiwada]

References:

1. Ashfold, T. (2014) *Understanding and Sourcing Political Opinion Polls.*
2. BBC News (2018) *The myth of the 2017 'youthquake' election.* Available at: https://www.bbc.co.uk/news/uk-politics-42747342 (Accessed: 05/06/24).
3. Laaser, U. *et al.* (2013) *Health Investigation: Analysis - Planning - Evaluation.*
4. National Association of Social Workers *Solving Ethical Dilemmas.* Available at: https://opentextbc.ca/ethicsinlawenforcement/chapter/solving-ethical-dilemmas/#return-footnote-823-1 (Accessed: 05/06/24).
5. Pickup, M. *et al.* (2011) 'Why Did the Polls Overestimate Liberal Democrat Support? Sources of Polling Error in the 2010 British General Election', *Journal of Elections, Public Opinion and Parties,* 21(2), pp. 179-209. Available at: https://doi.org/10.1080/17457289.2011.563309
6. Rogers, M. (2016) *Britain's EU referendum uncovers a key group – those who feel they have nothing to lose.* Available at: https://blogs.lse.ac.uk/government/2016/05/24/britains-eu-referendum-uncovers-a-key-group-those-who-feel-they-have-nothing-to-lose/
7. Sturgis, P. *et al.* (2017) 'An Assessment of the Causes of the Errors in the 2015 UK Gseneral Election Opinion Polls', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 181 Available at: https://doi.org/10.1111/rssa.12329