JUST IT

HANDBOOK

Data Technician bootcamp

(Data Search, Loading, Manipulation, Visualisation, Analysis)

- Excel
- Tableau
- Power BI
- SQL
- Programming language R

Student:
Alexei Kosyhin

10 Feb 2023

CONTENT

Page

Concepts

In this section I will outline everything what we learned about data.

- Data can be defined as an elementary value or the collection of values;
  Group Items which have subordinate data items are called Group item;
  Record can be defined as the collection of various data items;
  File can be defined as various records of one type of entity;
  Attribute and Entity represents the class of certain objects. It contains various attributes.

- There are many different sources of data. There are different types of files: MS Word Documents, Spreadsheets, Emails, PDFs, HTML, plaintext files, CSV, JSON,
- Defining Big Data – 4 bigs: volume, velocity, variety, veracity

- Open and Private Data

- Data Analysis Lifecycle: Gathering, Preparing, Choosing a model, Analysing, Presenting results, Making Decisions,

- Data Mapping Techniques: Define datatype and size; Map; Transform;
  Validate and Test; Deploy; Maintain and update.

- Data Structures: Relational Data Tables- fields, rows, values;
  Python – Strings, lists, tuples, sets, dictionaries, classes,
  functions …
- Structured and unstructured data

- Evolution to Big Data: Database servers -> Distributed data systems -> Onsite and Cloud computing Solutions

- Basic Data Management Technologies: Flat file database, Relational Database.

- Types of Data Analysis: Scalable Technologies, Business Intelligence.

- Conception of different types lists

- Purpose of Big Data Analysis: for Descriptive Analysis, Predictive Analysis, Prescriptive Analysis

- Types of Data: categorical data and numerical data. Qualitive categorical data can be nominal data and Ordinal data. Quantitative data can be Interval and Ration

-

Legislation relating to data security

Current Legislation: There are laws designed to protect users and their data from attack and misuse.

- **Computer Misuse Act 1990.** This includes planting a virus which is intended to cause damage, altering data, slowdown computer operation, frequent computer crashes. This particularly includes spreading malicious software like viruses, worms, Trojans, ..

- **Police and Justice Act 2006** (Computer Misuse) this is continuation of the Computer Misuse Act of 1990. This is particularly about hacking – unauthorised access to somebody computer.

- **The Copyright (Computer Programs) Regulation 1992**. From Wikipedia Software copyright is the application of copyright in law to machine-readable software.

- **Data Protection Acts (1998, 2018) and GDPR**. These acts are essentially abiut two points: The ethical use of personal data, and keeping individual’s personal data secure. Everyone responsible for using personal data has to follow strict rules called 'data protection principles'. They must make sure the information is: used fairly, lawfully and transparently.

- **Consumer Right Act 2015.** The aim of the 2015 Consumer Rights Act was to aid both consumers and retailers in understanding their rights and responsibilities, and thus to reduce and simplify disputes. It also seeks to encourage business based on fair practices and access to information.

Introduction to the problem – education via wealth and population

Even though the purpose of this handbook is to demonstrate pure technical skills, I will try to make my data manipulation, visualisation and analysis meaningful and interesting.

I decided to explore a topic about education in the world and how level of education (number of the best universities) depends on population and income. I found a list of 1000 best universities from Wikipedia even though total number of universities in the world is about 25,000. Then I found data of population and revenue by country from Wikipedia. Then I will try to make analysis and make conclusion.

I decided to use data from Wikipedia because Wikepedia is open source and doesn't require special permission for using data, Wikipedia is the most reliable and affordable source of good quality information.

Here are URLs for data from Wikipedia

1) List of top 1000 universities in the world - Wikipedia
2) List of countries by GDP (PPP) - Wikipedia
3) List of countries by population (United Nations) - Wikipedia

In order to provide analysis I will:

a) download the data using Power BI;

b) clean the data (delete empty rows, to replace wrong types e t c) using Power BI and Excel

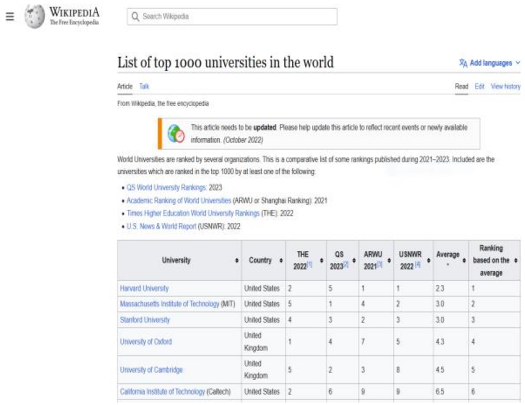
c) group by the list of universities using Power BI

d) visualize using Excel, Tableau, Power BI

e) merge/ join using Power BI, Microsoft SQL

f) draw some conclusion

Data manipulation – creating tables.



List of top 1000 universities in the world

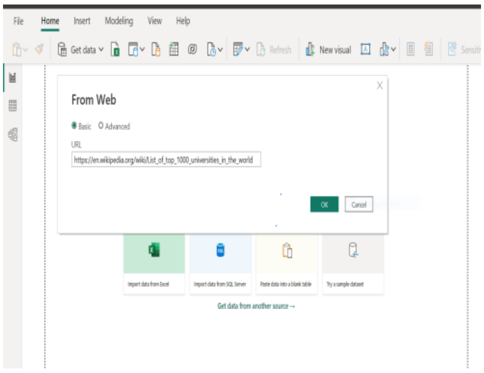
From Wikipedia, the free encyclopedia

This article needs to be **updated**. Please help update this article to reflect recent events or newly available information. *(October 2022)*

World Universities are ranked by several organizations. This is a comparative list of some rankings published during 2021–2023. Included are the universities which are ranked in the top 1000 by at least one of the following:

- QS World University Rankings: 2023
- Academic Ranking of World Universities (ARWU or Shanghai Ranking): 2021
- Times Higher Education World University Rankings (THE): 2022
- U.S. News & World Report (USNWR): 2022

University	Country	THE 2022[1]	QS 2023[2]	ARWU 2021[3]	USNWR 2022 [4]	Average *	Ranking based on the average
Harvard University	United States	2	5	1	1	2.3	1
Massachusetts Institute of Technology (MIT)	United States	5	1	4	2	3.0	2
Stanford University	United States	4	3	2	3	3.0	3
University of Oxford	United Kingdom	1	4	7	5	4.3	4
University of Cambridge	United Kingdom	5	2	3	8	4.5	5
California Institute of Technology (Caltech)	United States	2	6	9	9	6.5	6



From Web

● Basic ○ Advanced

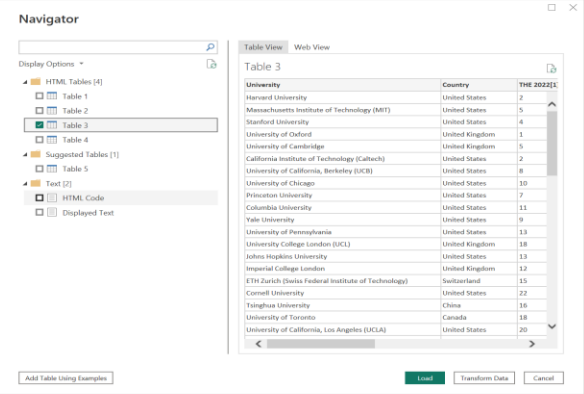
URL

https://en.wikipedia.org/wiki/List_of_top_1000_universities_in_the_world

OK Cancel

This pic. Shows original data in Wikipedia. This table consists 1000 rows.

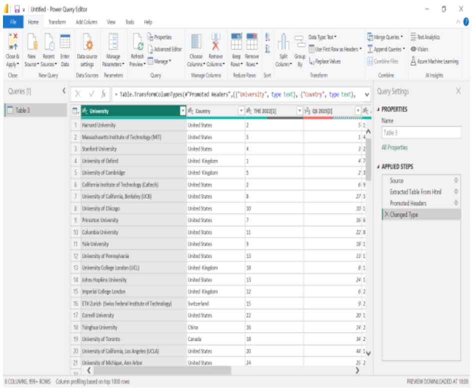
I download this table using Power BI



Navigator

Table 3

University	Country	THE 2022[1
Harvard University	United States	2
Massachusetts Institute of Technology (MIT)	United States	5
Stanford University	United States	4
University of Oxford	United Kingdom	1
University of Cambridge	United Kingdom	5
California Institute of Technology (Caltech)	United States	2
University of California, Berkeley (UCB)	United States	8
University of Chicago	United States	10
Princeton University	United States	7
Columbia University	United States	11
Yale University	United States	9
University of Pennsylvania	United States	13
University College London (UCL)	United Kingdom	18
Johns Hopkins University	United States	13
Imperial College London	United Kingdom	12
ETH Zurich (Swiss Federal Institute of Technology)	Switzerland	15
Cornell University	United States	22
Tsinghua University	China	16
University of Toronto	Canada	18
University of California, Los Angeles (UCLA)	United States	20



= Table.TransformColumnTypes(#"Promoted Headers",{{"University", type text}, {"Country", type text},

	University	Country	THE 2022[1]	QS 2023[2]
1	Harvard University	United States	2	5
2	Massachusetts Institute of Technology (MIT)	United States	5	1
3	Stanford University	United States	4	3
4	University of Oxford	United Kingdom	1	4
5	University of Cambridge	United Kingdom	5	2
6	California Institute of Technology (Caltech)	United States	2	6
7	University of California, Berkeley (UCB)	United States	8	27
8	University of Chicago	United States	10	10
9	Princeton University	United States	7	16
10	Columbia University	United States	11	22
11	Yale University	United States	9	18
12	University of Pennsylvania	United States	13	13
13	University College London (UCL)	United Kingdom	18	8
14	Johns Hopkins University	United States	13	24
15	Imperial College London	United Kingdom	12	6
16	ETH Zurich (Swiss Federal Institute of Technology)	Switzerland	15	9
17	Cornell University	United States	22	20
18	Tsinghua University	China	16	14
19	University of Toronto	Canada	18	34
20	University of California, Los Angeles (UCLA)	United States	20	44
21	University of Michigan, Ann Arbor	United States	24	25

PROPERTIES

Name

Table 3

All Properties

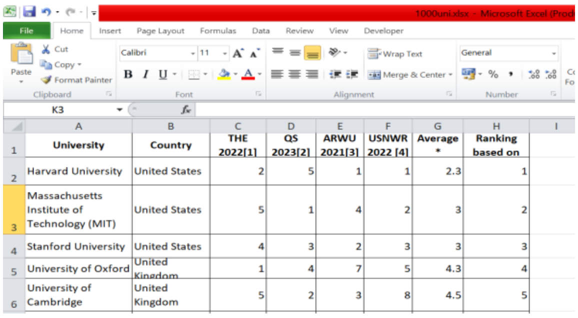
APPLIED STEPS

- Source
- Extracted Table From Html
- Promoted Headers
- Changed Type

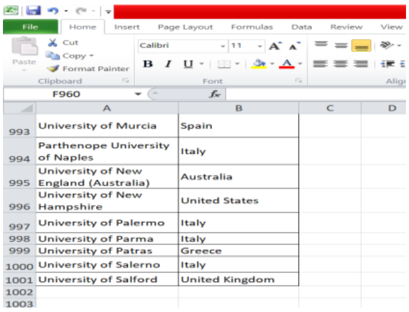
8 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 18:00

When I tried to load the table an error appeared. So, instead of load and used transform data. Then I copied the entire table and pasted into excel file.

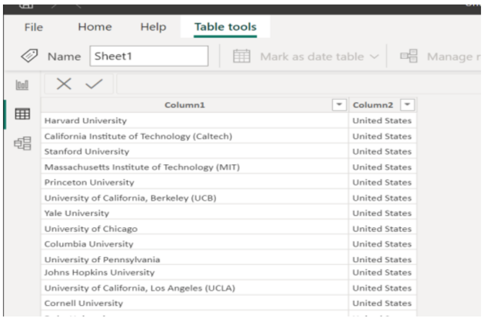


	A	B	C	D	E	F	G	H
1	University	Country	THE 2022[1]	QS 2023[2]	ARWU 2021[3]	USNWR 2022 [4]	Average *	Ranking based on
2	Harvard University	United States	2	5	1	1	2.3	1
3	Massachusetts Institute of Technology (MIT)	United States	5	1	4	2	3	2
4	Stanford University	United States	4	3	2	3	3	3
5	University of Oxford	United Kingdom	1	4	7	5	4.3	4
6	University of Cambridge	United Kingdom	5	2	3	8	4.5	5



	A	B	C	D
993	University of Murcia	Spain		
994	Parthenope University of Naples	Italy		
995	University of New England (Australia)	Australia		
996	University of New Hampshire	United States		
997	University of Palermo	Italy		
998	University of Parma	Italy		
999	University of Patras	Greece		
1000	University of Salerno	Italy		
1001	University of Salford	United Kingdom		
1002				
1003				

Once I load the table into excell, I simplified the data. I deleted some columns and left only 2. The table was big, has 1000 rows. I cleaned the table manually. Then I imported the data back to Power BI. The software didn't show an error.

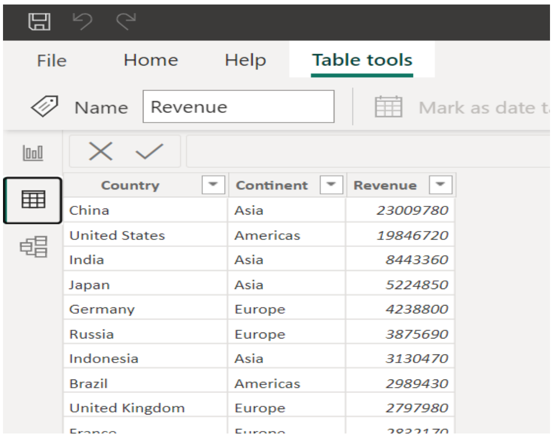


Name: Sheet1

Column1	Column2
Harvard University	United States
California Institute of Technology (Caltech)	United States
Stanford University	United States
Massachusetts Institute of Technology (MIT)	United States
Princeton University	United States
University of California, Berkeley (UCB)	United States
Yale University	United States
University of Chicago	United States
Columbia University	United States
University of Pennsylvania	United States
Johns Hopkins University	United States
University of California, Los Angeles (UCLA)	United States
Cornell University	United States

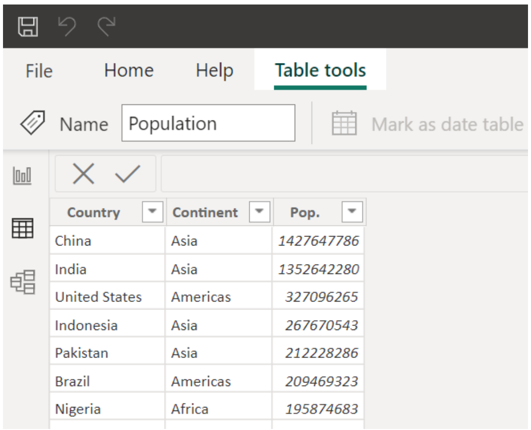
==

In the similar way I created in Power BI 2 tables – population by country and revenue by country. Now I have 3 tables in Power BI file. The data is cleaned enough for manipulation and analusis.



Name: Revenue

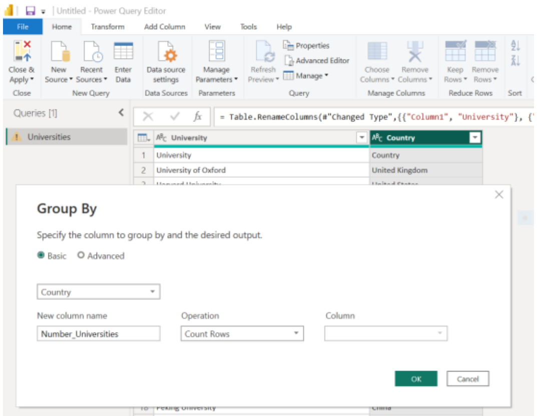
Country	Continent	Revenue
China	Asia	23009780
United States	Americas	19846720
India	Asia	8443360
Japan	Asia	5224850
Germany	Europe	4238800
Russia	Europe	3875690
Indonesia	Asia	3130470
Brazil	Americas	2989430
United Kingdom	Europe	2797980



Name: Population

Country	Continent	Pop.
China	Asia	1427647786
India	Asia	1352642280
United States	Americas	327096265
Indonesia	Asia	267670543
Pakistan	Asia	212228286
Brazil	Americas	209469323
Nigeria	Africa	195874683

In order to compare some metrics such as education /number of universities, revenue, population by country we need to transform some tables. For the table of list of universities we need to use "group by" function in Power BI to find number universities in each country.



```
= Table.RenameColumns(#"Changed Type",{{"Column1", "University"}, {
```

	University	Country
1	University	Country
2	University of Oxford	United Kingdom

Group By

Specify the column to group by and the desired output.

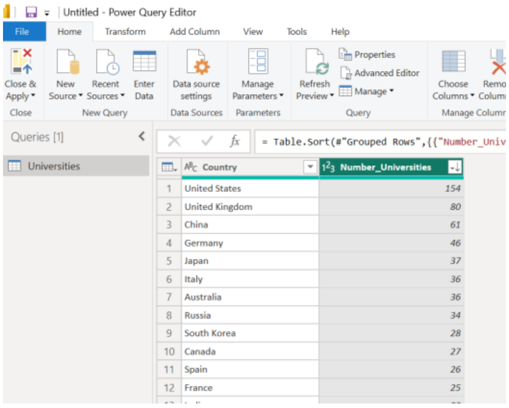
● Basic ○ Advanced

Country

New column name: Number_Universities

Operation: Count Rows

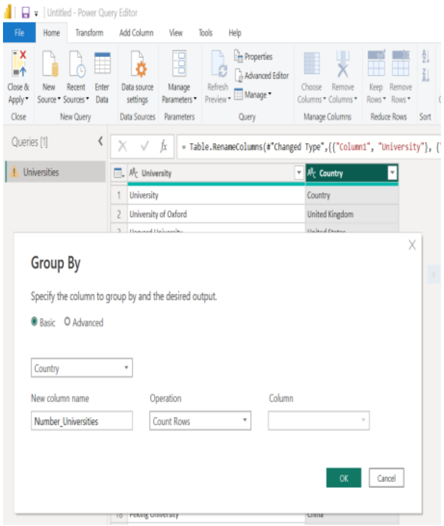
Column:



```
= Table.Sort(#"Grouped Rows",{{"Number_Univ
```

	Country	Number_Universities
1	United States	154
2	United Kingdom	80
3	China	61
4	Germany	46
5	Japan	37
6	Italy	36
7	Australia	36
8	Russia	34
9	South Korea	28
10	Canada	27
11	Spain	26
12	France	25

Using JOIN/ MERGE/ COMBINE in Power BI. For analysis and visualisation it would be nice to merge these 3 tables into one, key index will be “country”.



```
= Table.RenameColumns(#"Changed Type",{{"Column1", "University"}, {
```

	University	Country
1	University	Country
2	University of Oxford	United Kingdom

Group By

Specify the column to group by and the desired output.

● Basic ○ Advanced

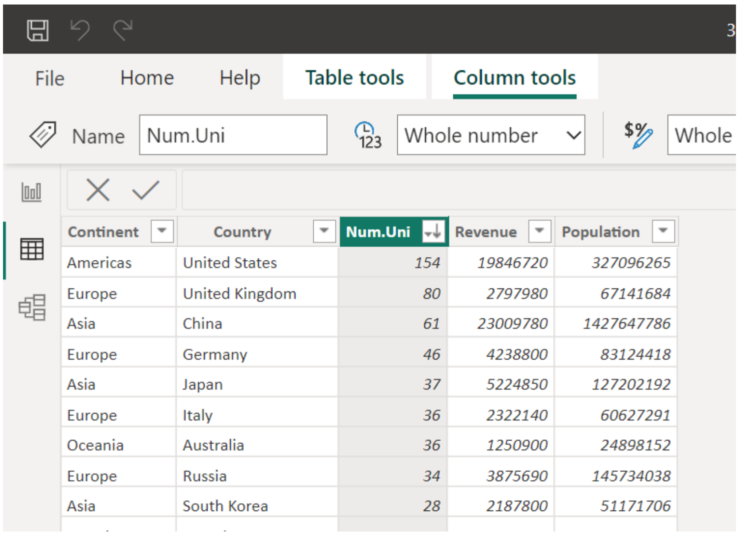
Country

New column name: Number_Universities

Operation: Count Rows

Column:

OK Cancel



Name: Num.Uni

Whole number

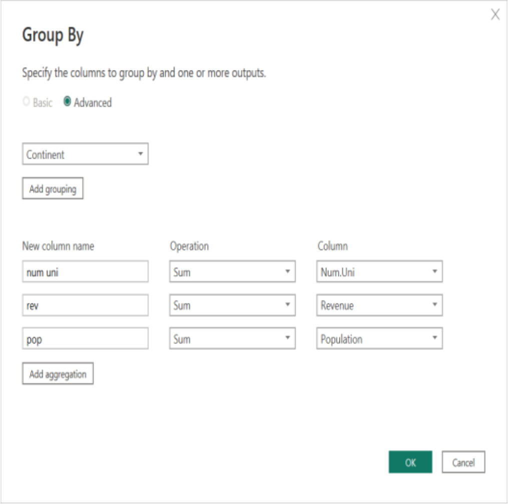
Whole

Continent	Country	Num.Uni	Revenue	Population
Americas	United States	154	19846720	327096265
Europe	United Kingdom	80	2797980	67141684
Asia	China	61	23009780	1427647786
Europe	Germany	46	4238800	83124418
Asia	Japan	37	5224850	127202192
Europe	Italy	36	2322140	60627291
Oceania	Australia	36	1250900	24898152
Europe	Russia	34	3875690	145734038
Asia	South Korea	28	2187800	51171706

Here I used Power BI Query functions
To combine 3 tables into one.

This is the main table from which we will do visualisation and analysis.

However a manager was curious about the data for each continent. For this purpose I will group the table by key “Continents”.



Group By

Specify the columns to group by and one or more outputs.

○ Basic ● Advanced

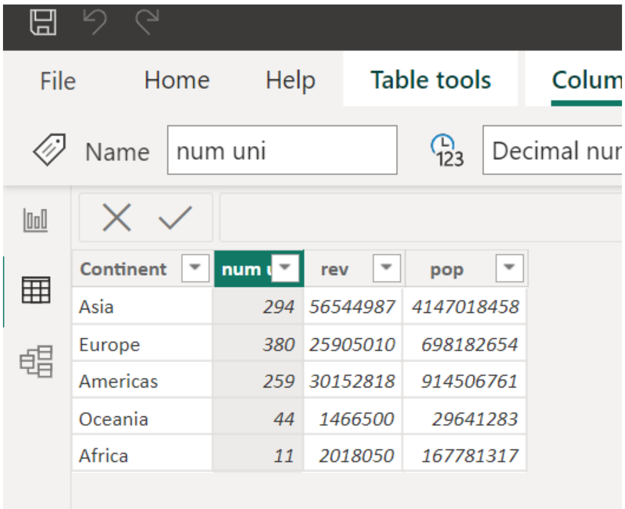
Continent

Add grouping

New column name	Operation	Column
num uni	Sum	Num.Uni
rev	Sum	Revenue
pop	Sum	Population

Add aggregation

OK Cancel

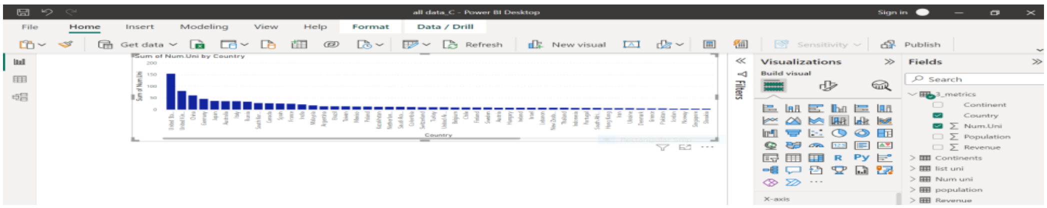


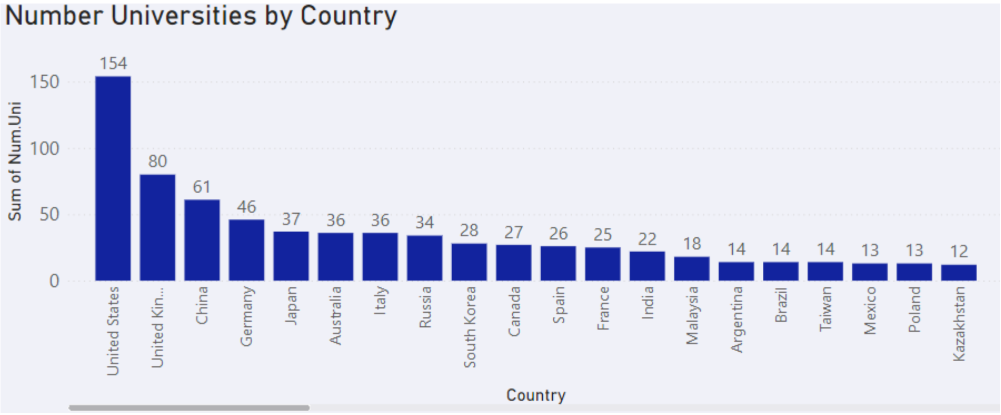
Name: num uni

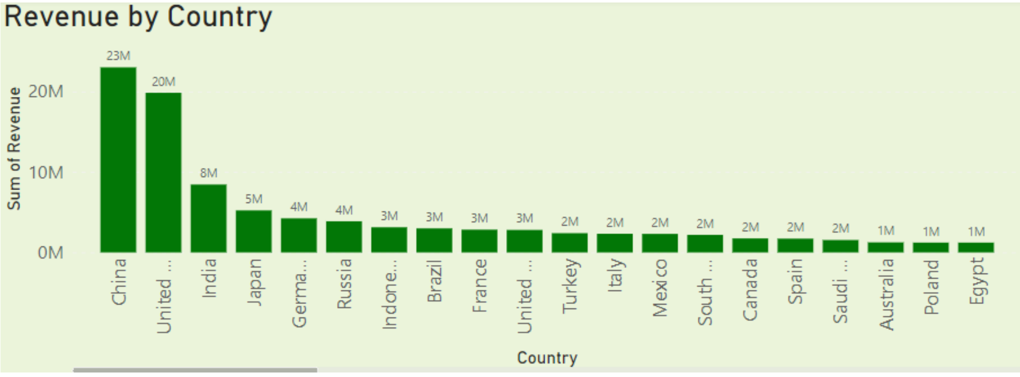
Decimal nur

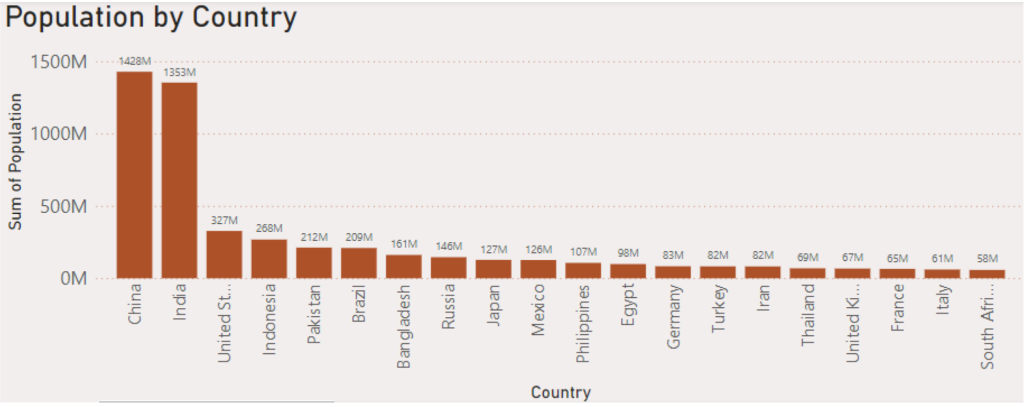
Continent	num u	rev	pop
Asia	294	56544987	4147018458
Europe	380	25905010	698182654
Americas	259	30152818	914506761
Oceania	44	1466500	29641283
Africa	11	2018050	167781317

VISUALIZATION using Power BI tools



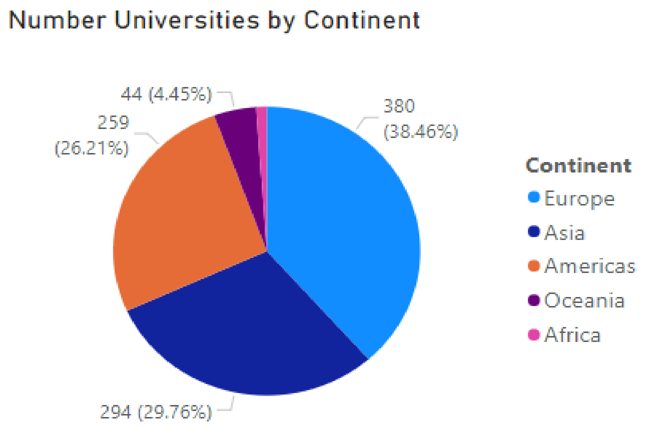


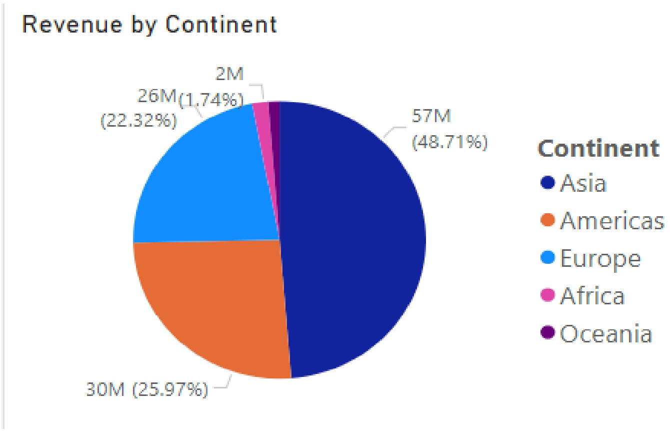


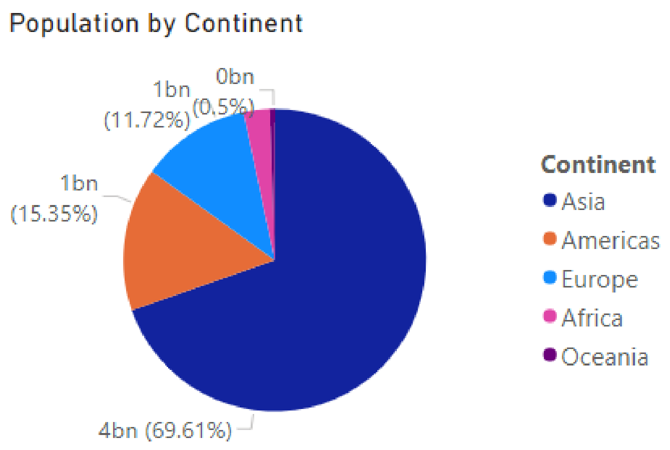


Visualisation using Power Bi

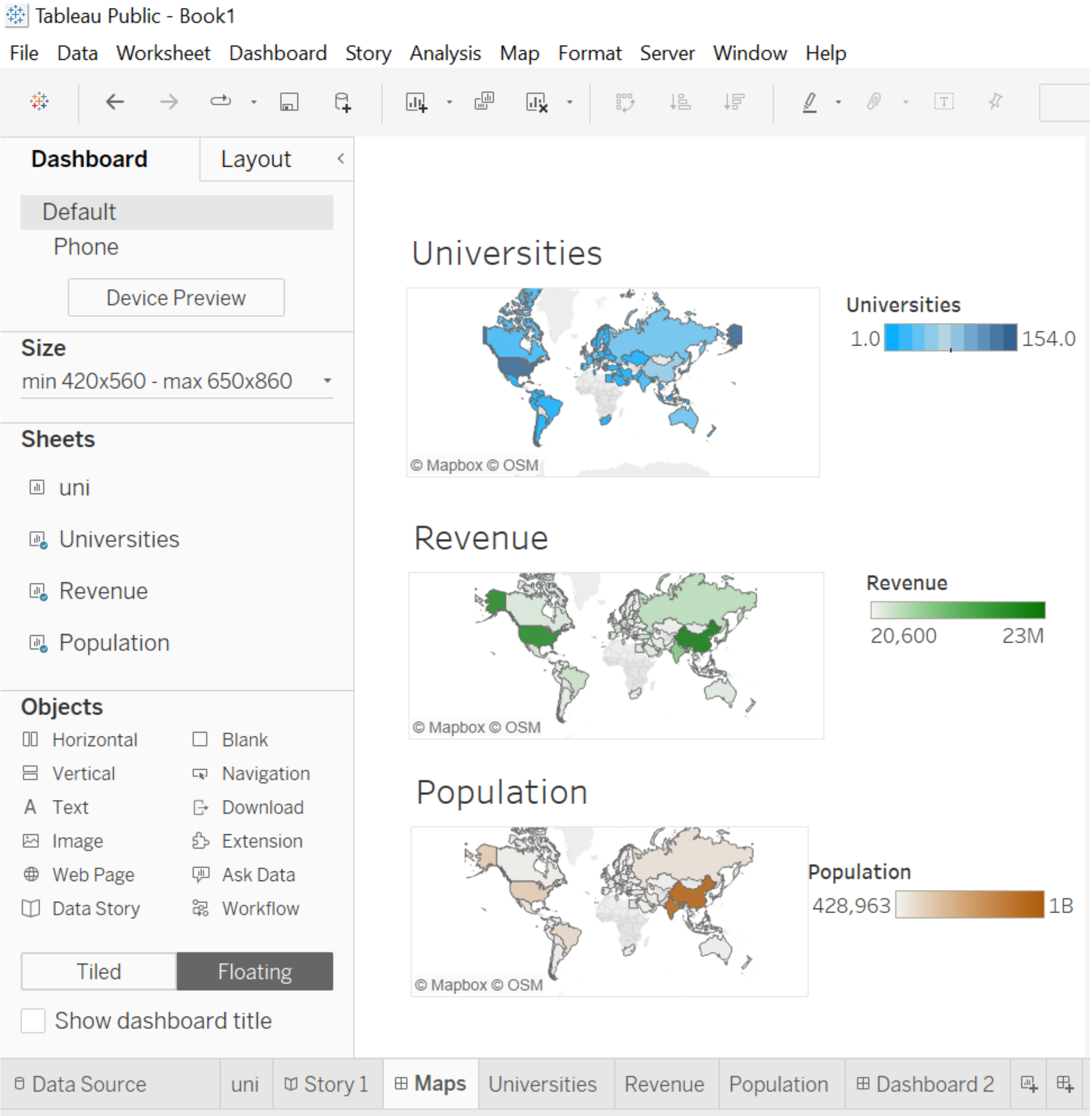
(Displaying numbers of universities, population and revenue by continent)



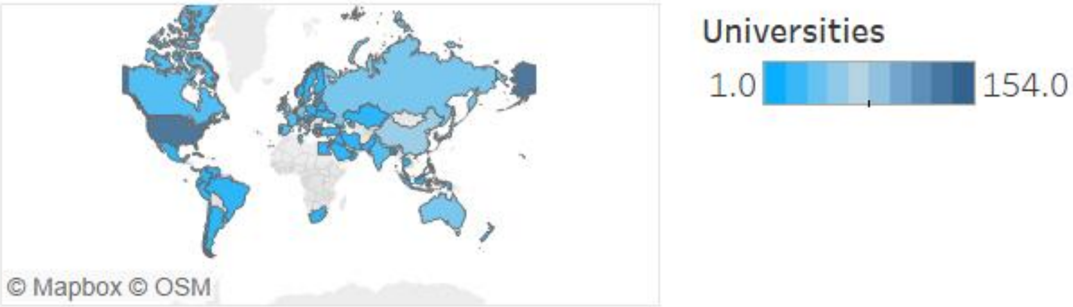




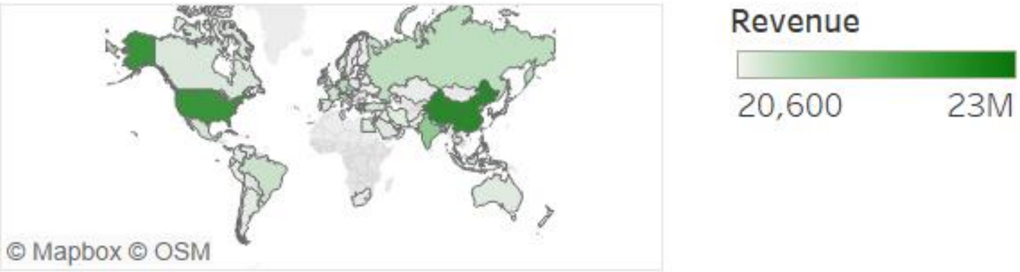
Visualization using TABLEAU.



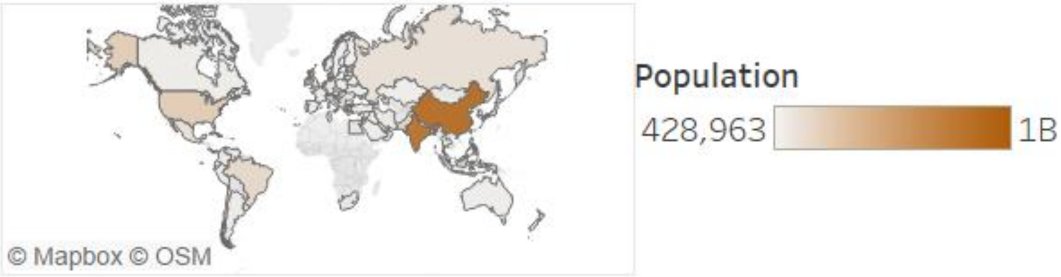
Universities



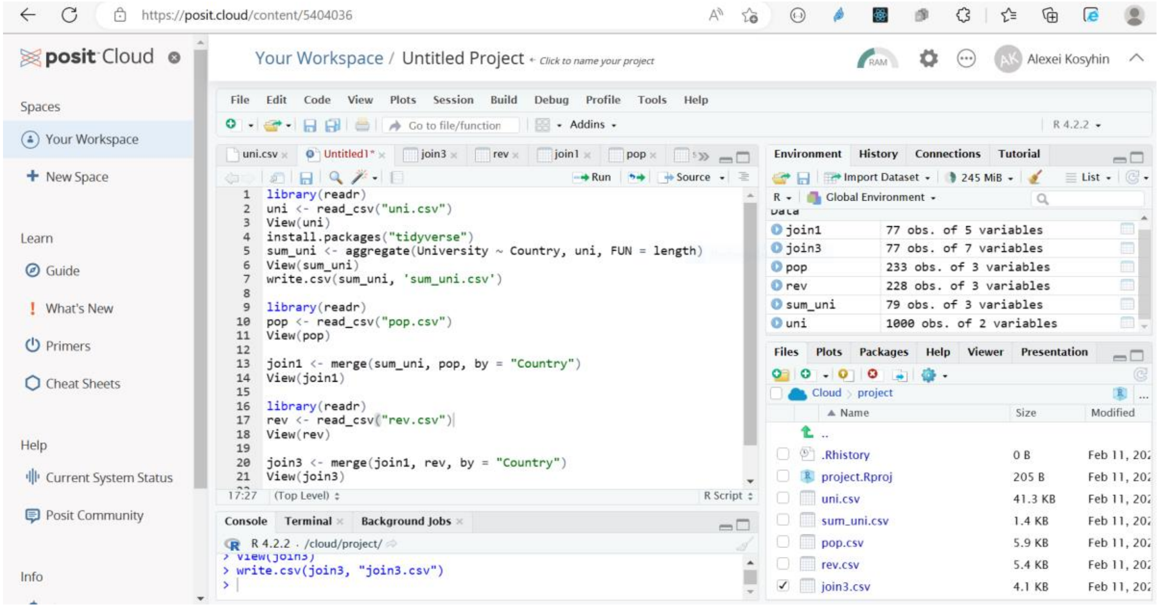
Revenue



Population



Data Manipulation in R Studio


https://posit.cloud/content/5404036
posit Cloud
Spaces
Your Workspace
New Space
Learn
Guide
What's New
Primers
Cheat Sheets
Help
Current System Status
Posit Community
Info
Your Workspace / Untitled Project ← Click to name your project
RAM
Alexei Kosyhin
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.2
uni.csv
Untitled1*
join3
rev
join1
pop
Run
Source
Environment History Connections Tutorial
Import Dataset
245 MiB
List
R
Global Environment
Data
Files Plots Packages Help Viewer Presentation
Cloud › project
Console Terminal Background Jobs

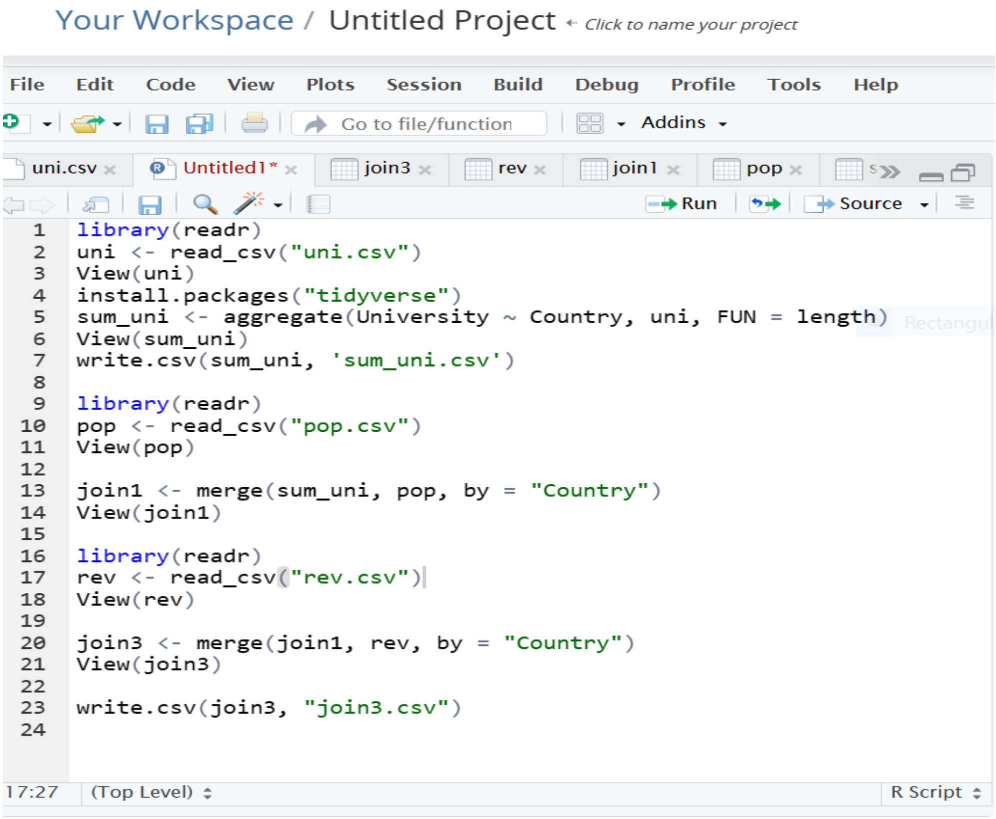
```r
library(readr)
uni <- read_csv("uni.csv")
View(uni)
install.packages("tidyverse")
sum_uni <- aggregate(University ~ Country, uni, FUN = length)
View(sum_uni)
write.csv(sum_uni, 'sum_uni.csv')

library(readr)
pop <- read_csv("pop.csv")
View(pop)

join1 <- merge(sum_uni, pop, by = "Country")
View(join1)

library(readr)
rev <- read_csv("rev.csv")
View(rev)

join3 <- merge(join1, rev, by = "Country")
View(join3)
```

17:27 (Top Level) ÷ R Script ÷

join1	77 obs. of 5 variables
join3	77 obs. of 7 variables
pop	233 obs. of 3 variables
rev	228 obs. of 3 variables
sum_uni	79 obs. of 3 variables
uni	1000 obs. of 2 variables

	Name	Size	Modified
	..		
□	.Rhistory	0 B	Feb 11, 20:
□	project.Rproj	205 B	Feb 11, 20:
□	uni.csv	41.3 KB	Feb 11, 20:
□	sum_uni.csv	1.4 KB	Feb 11, 20:
□	pop.csv	5.9 KB	Feb 11, 20:
□	rev.csv	5.4 KB	Feb 11, 20:
☑	join3.csv	4.1 KB	Feb 11, 20:

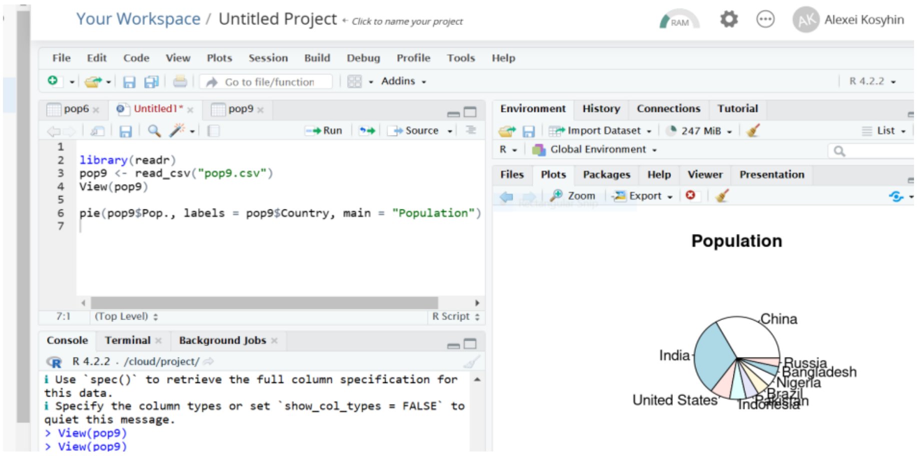
```
R 4.2.2 · /cloud/project/
> View(join3)
> write.csv(join3, "join3.csv")
> 
```


Your Workspace / Untitled Project ← Click to name your project
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
uni.csv
Untitled1*
join3
rev
join1
pop
Run
Source

```r
library(readr)
uni <- read_csv("uni.csv")
View(uni)
install.packages("tidyverse")
sum_uni <- aggregate(University ~ Country, uni, FUN = length)
View(sum_uni)
write.csv(sum_uni, 'sum_uni.csv')

library(readr)
pop <- read_csv("pop.csv")
View(pop)

join1 <- merge(sum_uni, pop, by = "Country")
View(join1)

library(readr)
rev <- read_csv("rev.csv")
View(rev)

join3 <- merge(join1, rev, by = "Country")
View(join3)

write.csv(join3, "join3.csv")

```

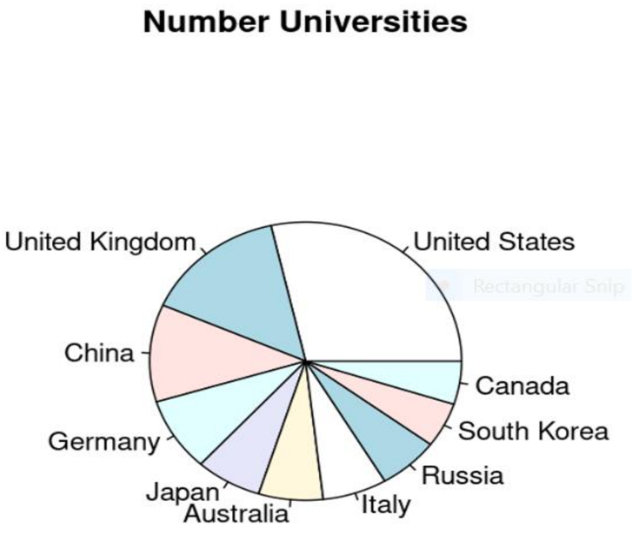
17:27 (Top Level) ÷ R Script ÷

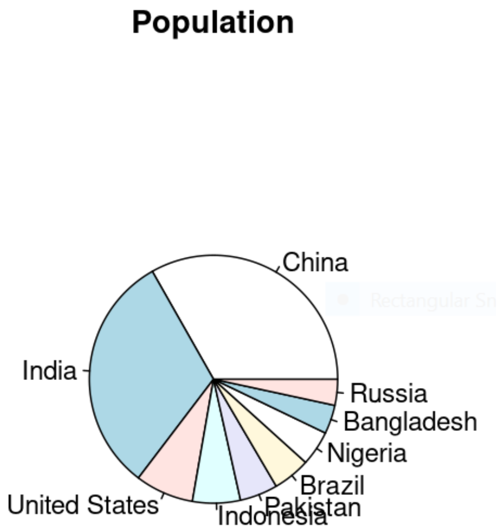
Visualization in R Studio

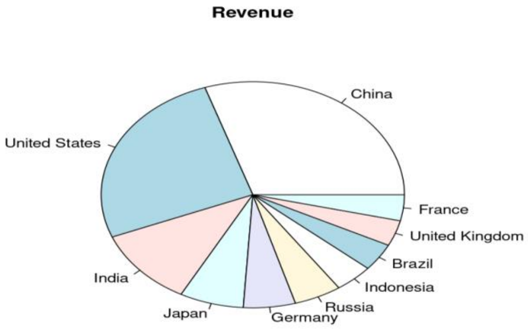

Your Workspace / Untitled Project ← Click to name your project
RAM
Alexei Kosyhin
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.2.2
pop6
Untitled1*
pop9
Run
Source
Environment History Connections Tutorial
Import Dataset
247 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Zoom
Export
7:1
(Top Level)
R Script
Console Terminal Background Jobs
R 4.2.2 · /cloud/project/
Population
China
India
Russia
Bangladesh
Nigeria
Brazil
Pakistan
Indonesia
United States

```
library(readr)
pop9 <- read_csv("pop9.csv")
View(pop9)

pie(pop9$Pop., labels = pop9$Country, main = "Population")
```

```
i Use `spec()` to retrieve the full column specification for
this data.
i Specify the column types or set `show_col_types = FALSE` to
quiet this message.
> View(pop9)
> View(pop9)
```







Analysis/Conclusion

Here I will provide a simple analysis how level education/ number of top universities in each country depends on revenue and population.

From above graphs we can see an obvious correlations 1) between number of universities and revenue, and 2) between number of universities and population.

In general, the more country's avenue or population the more universities in the country.

For example, the USA is the richest country and has the biggest number top universities.

For example, India is the most populated country and has the significant number of universities.

There are some exceptions, for example African countries have a very high population, but only few universities.

However other factors should be considered for example an official country language.
I would speculate that the UK has a great number of universities particularly because of English language (everybody wants to learn English in an English speaking country)

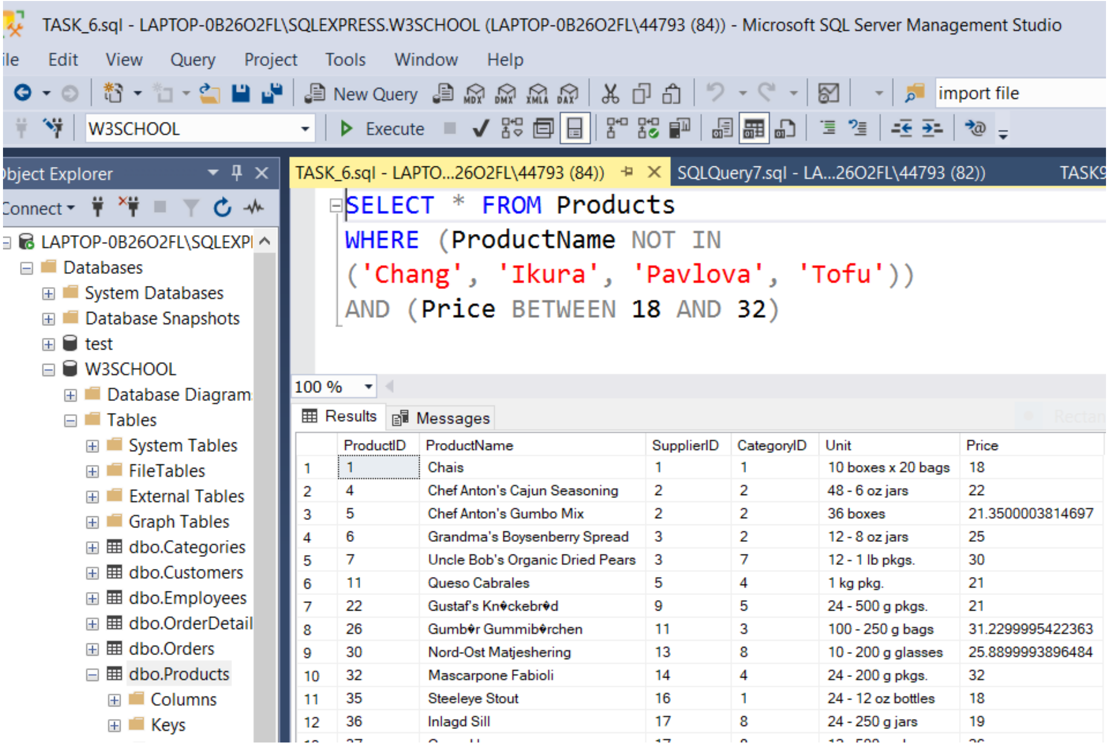
SQL – Big questions

Task 6

The manager of Northwind Company Limited has marked out certain products for promotional discounts and these would need to be offered as Christmas Discounted Price.

Please create a report containing all products apart from chang, Ikura, Pavlova, and tofu. Ensuring that they relate products with the price ranging from 18 to 32

Please sort the Christmas Discounted Price in descending order



```
SELECT * FROM Products
WHERE (ProductName NOT IN
('Chang', 'Ikura', 'Pavlova', 'Tofu'))
AND (Price BETWEEN 18 AND 32)
```

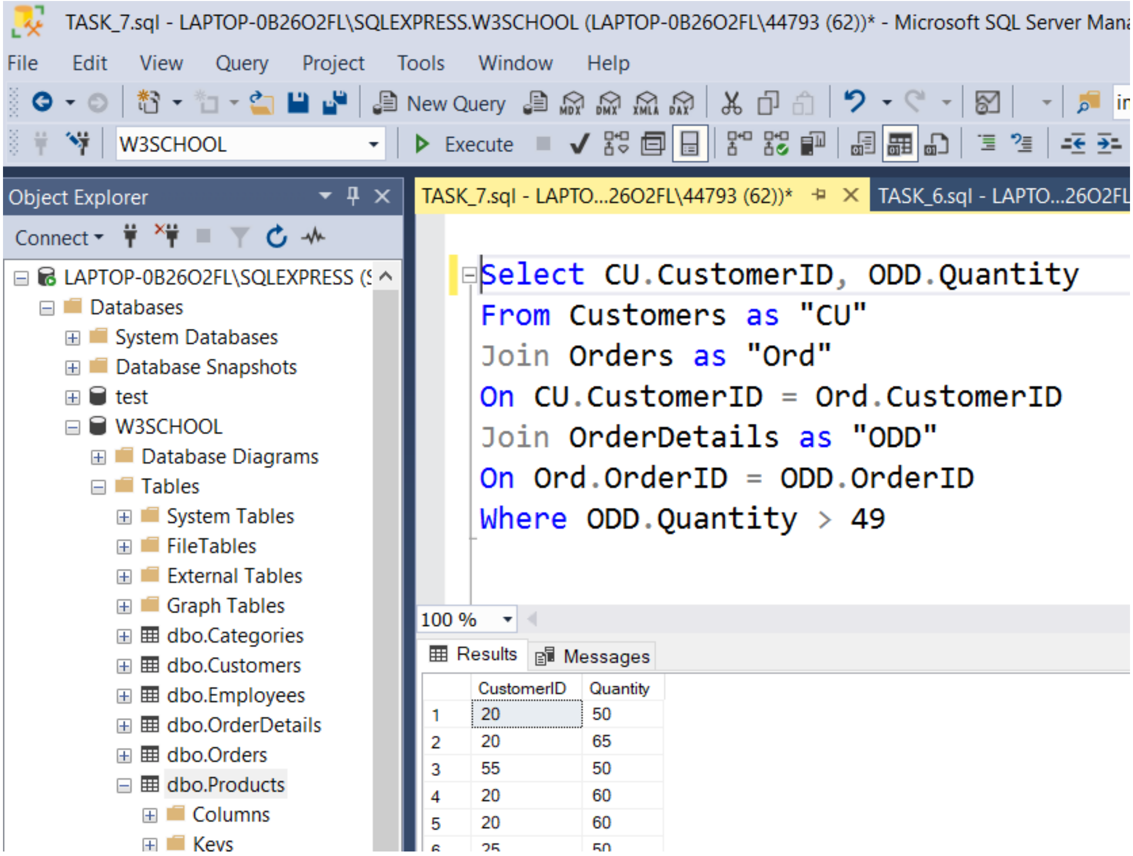
	ProductID	ProductName	SupplierID	CategoryID	Unit	Price
1	1	Chais	1	1	10 boxes x 20 bags	18
2	4	Chef Anton's Cajun Seasoning	2	2	48 - 6 oz jars	22
3	5	Chef Anton's Gumbo Mix	2	2	36 boxes	21.3500003814697
4	6	Grandma's Boysenberry Spread	3	2	12 - 8 oz jars	25
5	7	Uncle Bob's Organic Dried Pears	3	7	12 - 1 lb pkgs.	30
6	11	Queso Cabrales	5	4	1 kg pkg.	21
7	22	Gustaf's Kn�ckebr�d	9	5	24 - 500 g pkgs.	21
8	26	Gumb�r Gummib�rchen	11	3	100 - 250 g bags	31.2299995422363
9	30	Nord-Ost Matjeshering	13	8	10 - 200 g glasses	25.8899993896484
10	32	Mascarpone Fabioli	14	4	24 - 200 g pkgs.	32
11	35	Steeleye Stout	16	1	24 - 12 oz bottles	18
12	36	Inlagd Sill	17	8	24 - 250 g jars	19

```
SELECT * FROM [Products]
WHERE (ProductName NOT IN ("Chang", "Ikura", "Pavlova", "Tofu"))
AND (PRICE BETWEEN 18 AND 32);
```

Task 7

The manager of Northwind Company Limited wants to embark on promotional activities to focus on their most active customers.

She has therefore asked me to prepare a report that will return to her, the most active customers who have order for goods with quantities that are 50 or above, as well as the quantity ordered



```
Select CU.CustomerID, ODD.Quantity
From Customers as "CU"
Join Orders as "Ord"
On CU.CustomerID = Ord.CustomerID
Join OrderDetails as "ODD"
On Ord.OrderID = ODD.OrderID
Where ODD.Quantity > 49
```

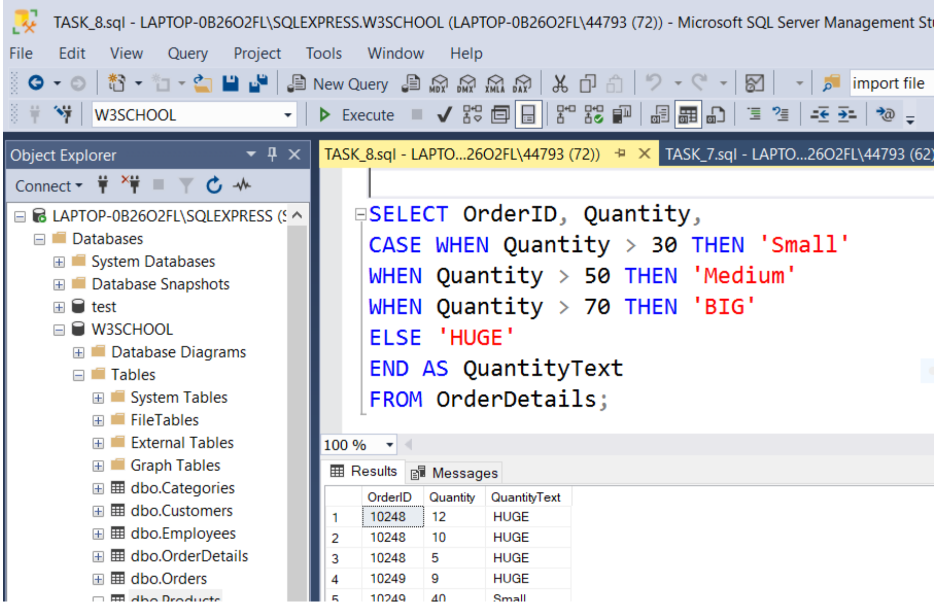
	CustomerID	Quantity
1	20	50
2	20	65
3	55	50
4	20	60
5	20	60
6	25	50

Select CU.CustomerID, ODD.Quantity
From Customers as "CU"
Join Orders as "Ord"
On CU.CustomerID = Ord.CustomerID
Join OrderDetails as "ODD"
On Ord.OrderID = ODD.OrderID
Where ODD.Quantity > 49

Task 8

Return to me all orders that have been made at Northwind Company Limited

Please grade them according to the size of quantity ordered e.g. Low value customer to refer to those who order for quantities that 30 or less; Value customer to refer to those who order for quantities that 50 or less; Large customer to refer to those who order for quantities that 70 or less; and finally Premium customers for the others



```sql
SELECT OrderID, Quantity,
CASE WHEN Quantity > 30 THEN 'Small'
WHEN Quantity > 50 THEN 'Medium'
WHEN Quantity > 70 THEN 'BIG'
ELSE 'HUGE'
END AS QuantityText
FROM OrderDetails;
```

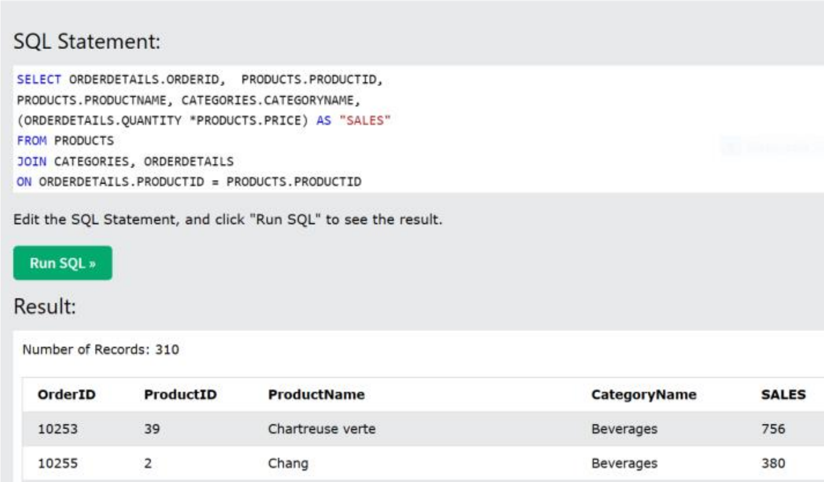
	OrderID	Quantity	QuantityText
1	10248	12	HUGE
2	10248	10	HUGE
3	10248	5	HUGE
4	10249	9	HUGE
5	10249	40	Small

```
SELECT OrderID, Quantity,
CASE WHEN Quantity  < 30 THEN 'SMALL'
WHEN Quantity < 50 THEN "NORMAL"
WHEN Quantity < 80 THEN "BIG"
ELSE 'EXTRA LARGE'
END AS QuantityText FROM OrderDetails;
```

Task 9

Return to me the sales of Northwind Company Limited according to their categories.

Please ensure that your report only concentrates on Sales that are in excess of 300



SQL Statement:

```
SELECT ORDERDETAILS.ORDERID,  PRODUCTS.PRODUCTID,
PRODUCTS.PRODUCTNAME, CATEGORIES.CATEGORYNAME,
(ORDERDETAILS.QUANTITY *PRODUCTS.PRICE) AS "SALES"
FROM PRODUCTS
JOIN CATEGORIES, ORDERDETAILS
ON ORDERDETAILS.PRODUCTID = PRODUCTS.PRODUCTID
```

Edit the SQL Statement, and click "Run SQL" to see the result.

Result:

Number of Records: 310

OrderID	ProductID	ProductName	CategoryName	SALES
10253	39	Chartreuse verte	Beverages	756
10255	2	Chang	Beverages	380

```
SELECT ORDERDETAILS.ORDERID,  PRODUCTS.PRODUCTID,
PRODUCTS.PRODUCTNAME, CATEGORIES.CATEGORYNAME,
(ORDERDETAILS.QUANTITY *PRODUCTS.PRICE) AS "SALES"
FROM PRODUCTS
JOIN CATEGORIES, ORDERDETAILS
ON ORDERDETAILS.PRODUCTID = PRODUCTS.PRODUCTID

AND CATEGORIES.CATEGORYID = PRODUCTS.CATEGORYID
WHERE SALES > 300
ORDER BY CATEGORYNAME
```

Question 10

To specify multiple values (in order words, a shorthand for multiple OR conditions), which of the following Operators is can be used

a. IN

b. LIKE

c. BETWEEN

d. AND

Question 11

When an aggregate SQL statement is involved, the SQL Clause which is used instead of a Where Clause is called

a. WHERE Clause

b. Group By Clause

c. Case Statement

d. The LIKE Condition

Question 12

The Keyword that eliminates duplicate rows from the results of a SELECT Statement is called

a. Select *

b. DISTINCT

c. Avoid Duplicate

d. The IN Operator

Question 13

The Clause which sorts data of a table is called

a. The Group By Clause

b. Having Clause

c. Order By Clause

d. The Where Clause

Question 14

We use the ……**LIKE**……… Operator with SELECT to set conditions based on pattern matching

a. IN

b. LIKE

c. BETWEEN

d. NOT