Data Mining & Big Data

Exercise 1 Data types

Give for each following variables its type and an example:

1. place of residence
2. citizenship
3. number of children
4. gender
5. matrimonial state
6. age
7. annual income
8. shoe size
9. education level
10. native language
11. number of language spoken
12. pants size
13. satisfaction

Exercise 2 Google trends

Google Trends is a web tool that presents how often a particular search-term is entered on Google search. The study was performed on January 2018, the 12th on the words *infogram*, *qlikview*, *plotly* and *tableau* in the category *computer science and electronics*. The following figures show *infogram* in blue, *qlikview* in red, *plotly* in yellow and *tableau* in green.

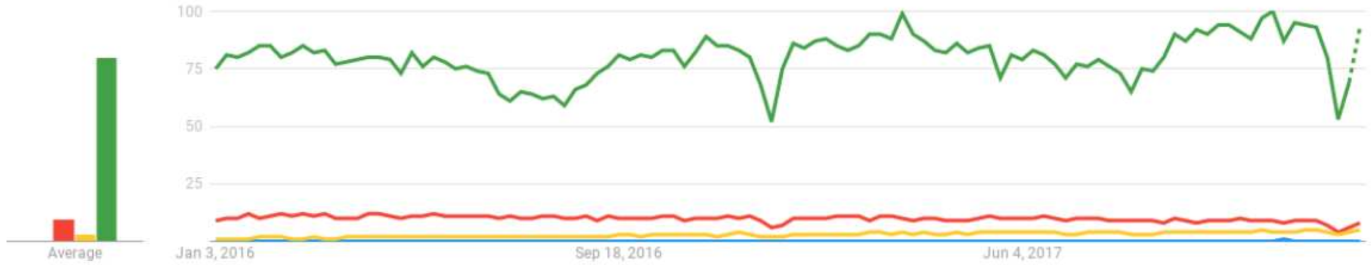
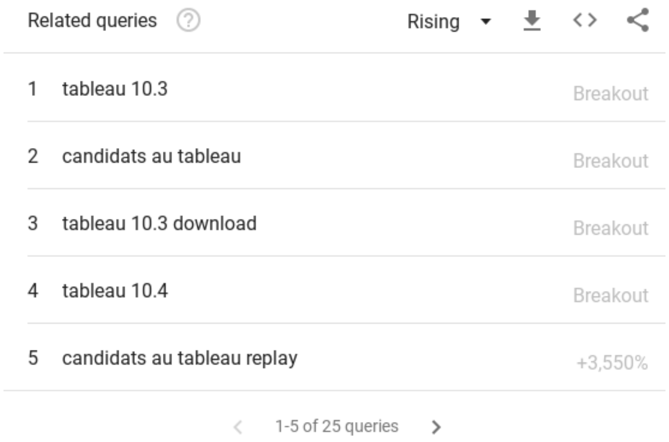


Figure 1: Line chart for infogram, qlikview, plotly and tableau.



1	tableau 10.3	Breakout
2	candidats au tableau	Breakout
3	tableau 10.3 download	Breakout
4	tableau 10.4	Breakout
5	candidats au tableau replay	+3,550%

Figure 2: Related queries for tableau.

1. Concerning figure 1, what is the period presented ? What information can you extract from the figure ?
2. Figure 2 presents the most popular queries associated to the word *tableau*. What can you conclude ?

We decide to delete the term *tableau* in order to visualize better other terms. The figures 3 and 4 correspond to the result obtained.

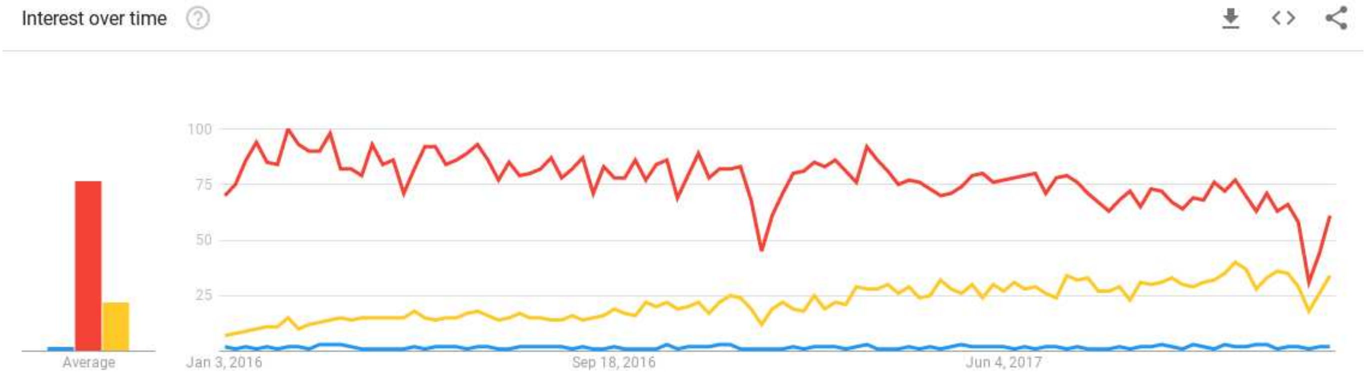


Figure 3: Line chart for infogram, qlikview and plotly.

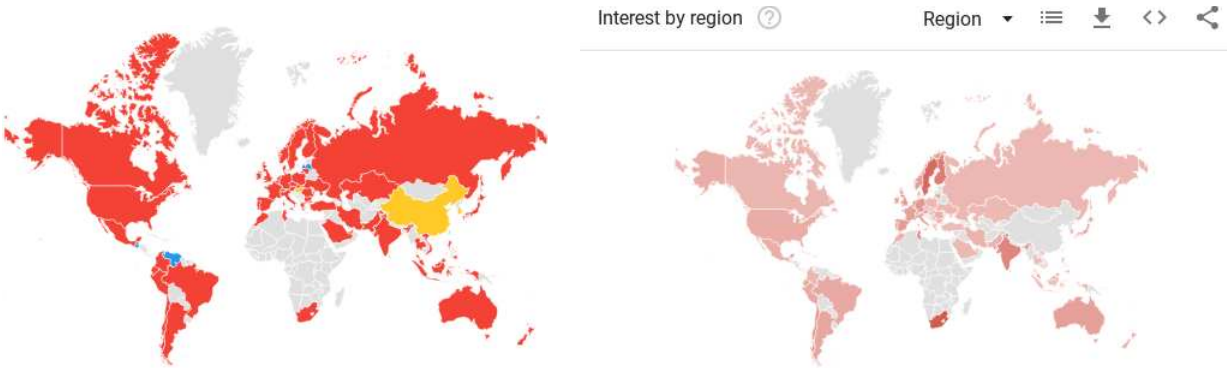
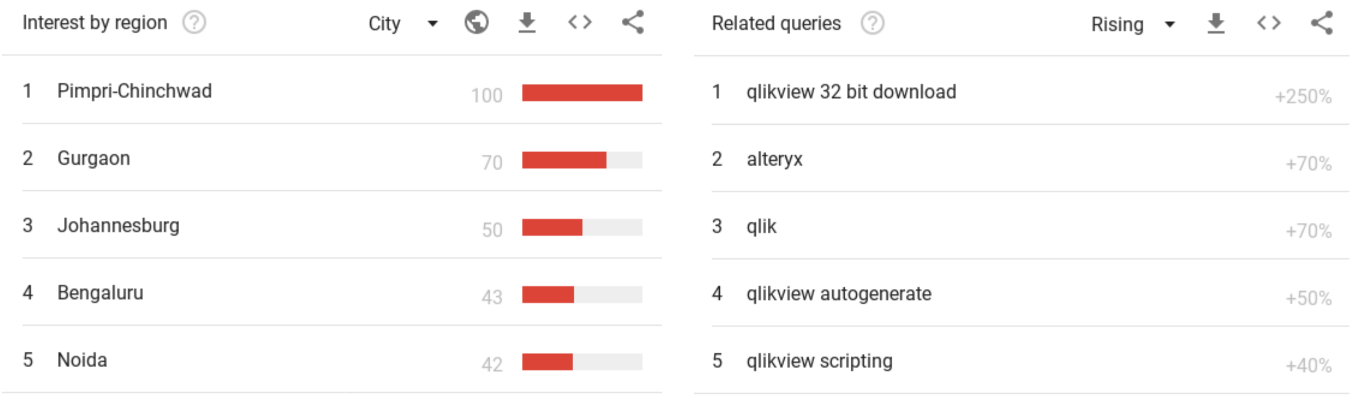


Figure 4: Map for all softwares then just for qlikview.



1	Pimpri-Chinchwad	100
2	Gurgaon	70
3	Johannesburg	50
4	Bengaluru	43
5	Noida	42

1	qlikview 32 bit download	+250%
2	alteryx	+70%
3	qlik	+70%
4	qlikview autogenerate	+50%
5	qlikview scripting	+40%

Figure 5: Cities and related queries for qlikview.

3. Analyse the tendancy of the three softwares with figure 1.
4. Enunciate one or two countries which use the most *infogram* for web research. Do the same for *qlikview* and *plotly.*
5. For *qlikview*, does the related queries corresponding to the sofware ? Are all the queries taking the same importance ?

Exercise 3 Mosaic plot

The following table of values shows a sample of 2300 music listeners classified by age, education and whether they listen to classical music.

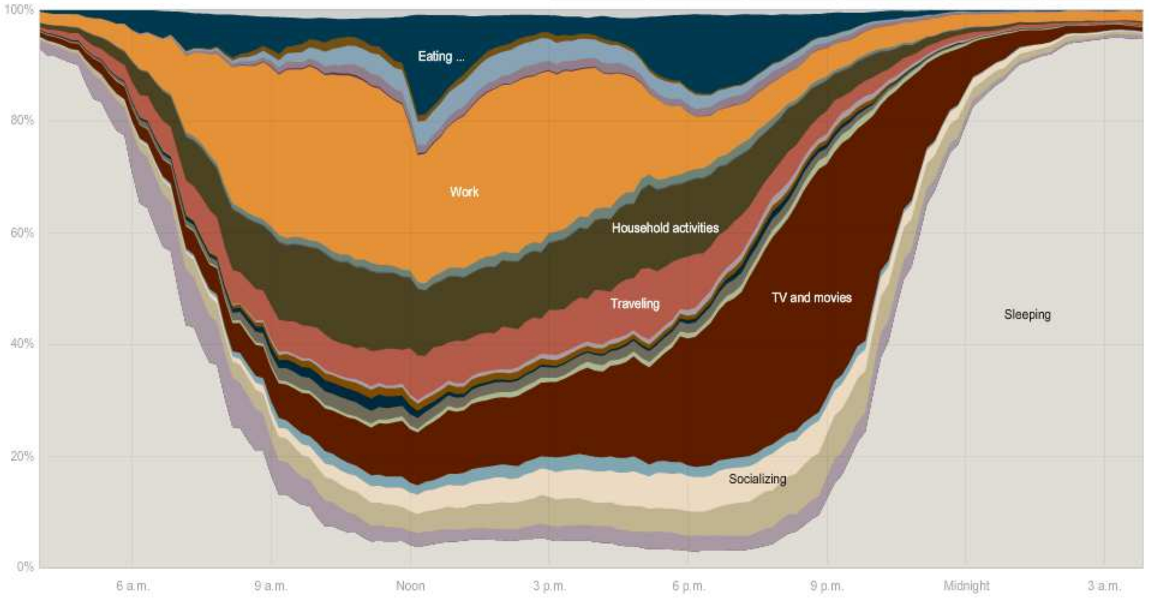
	Education high		Education low	
age \ classical music	yes	no	yes	no
old	210	190	170	730
young	194	406	110	290

1. Create a mosaic plot from the table. Use first the age, then the education level and finally the musical taste.
2. What can you observe ?

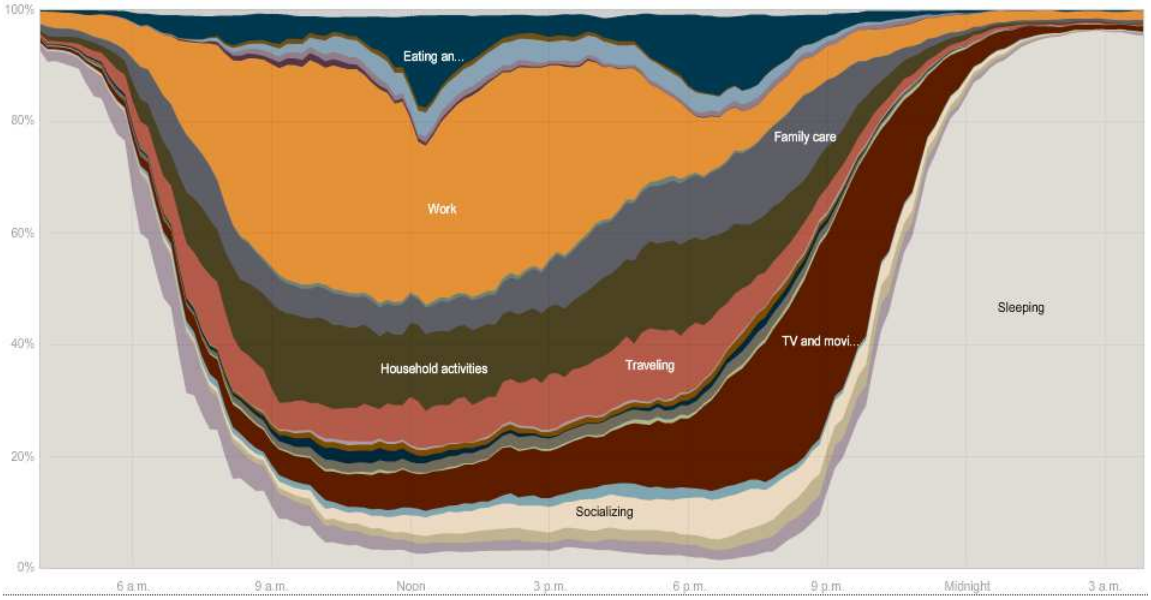
Exercise 4 Stream graph visualization

The following stream graph presents the results of an American Time Use Survey (2008) asked to thousands of American residents over age 15. They had to recall every minute of a day.

1. Compare activities of people living with no child vs activities for those with two or more children. What is different and what is not ?



no child



two or more children

www.nytimes.com/interactive/2009/07/31/business/20080801-metrics-graphic.html