Table of Contents

Introduction

Data Science has been regarded as a multi–disciplinary field in which the role of a so–called Data Scientist varies from one company to another and tends to overlap those of other data–related positions. Within the past years, the required set of Data Science skills has escalated greatly in such number and complexity that it becomes overwhelming and intimidating for novices at early stages of their career. This narrative visualisation aims to help aspiring beginners navigate their careers by providing realistic insights into Data Science job market. It seeks to answer 3 frequently asked questions by anyone wishing to break into the industry:

What am I expected do? → *What am I required to have?* → *Where can I look for a job?*

It first walks users through key responsibilities of a Data Scientist and later to job requirements, which are examined under 4 facets: Technical skills (programming languags and software proficiency), Domain skills (area of study e.g., Mathematics, Computer Science), Education level (Bachelor, Master, PhD) and Years of Experience. It also suggests a number of geographical regions and business industries with the most promising opportunities for Data Science jobs. Though Data Scientist remains at the heart of the narrative, the visualisation also compares and contrasts Data Scientist with Data Analyst, Machine Learning Engineer, Data Engineer, Big Data Developer and Analytics Consultant on the aforementioned aspects, and users are encouraged to consider these alternatives as part of their career planning. In addition, it supports the exploration of the variety across experience levels, from which not only entry level candidates can benefit but also those at the higher levels can find valuable insights as they move up the ladder.

Design

The design process begins by specifying the concepts to be visualised in order to answer the research questions. These concepts strongly align with the logic that aspiring novices have in mind when seeking a job, including key responsibilities, essential requirements and job demand for the position. Then to decide on the appropriateness of visual representation, one must consider the types of data corresponding to each concept. Data on job duties are purely textual, so popular text visualisations include word tree and word cloud. While word cloud is keyword–based, thereby much less informative and self–explanatory, word tree is based on sequences of words, so would be far more useful

and straightforward if the sequences are phrases describing job activities in the *"To do what"* structure. With its data also both textual and extracted from job descriptions, job requirement attribute nevertheless can be decomposed into skills (categorical data), education levels (categorical data) and years of experience (numeric data), which on the other hand provides more visualisation options. A skill can be a programming language, a tool or a software that a candidate is required to be proficient in, and the importance of a skill can be measured by its frequency in the document. Skills can be plotted against their frequencies using a dot plot or a lollipop chart, on which number of dots or length of a lollipop represents frequency counts. However, the list is made up of up to 130 items, and given such number of categories, it is aesthetically undesirable to have them all plotted in the same space. This makes word cloud an outstanding option in this case as these 130 words can be plotted neatly and visibly within a cloud with more popular words (i.e., more often required skills) poping up around the center in bigger sizes.

With regard to education, this analysis discusses minimum level of degree required by the employers and there are 3 distinct categories – Bachelor, Master, PhD level. A side–by–side column chart is perfectly suitable in this situation where the height of a bar indicates the number of hiring positions associated with a certain category. Similarly, experience is also examined under the concept of minimum years of experience required. It is necessary to first compare the proportion of jobs that did specify the number and those that did not, for which a simple stacked bar can do the work. Then it can be observed that the number of years of experience is correlated with education level. Intuitively, a hiring position for which a candidate is expected to have at least a Master degree is more likely to require higher number of years. Thus, a line chart is useful to account for this possibly upward trend. To avoid redundancy, the line chart is incorporated into the side–by–side column chart to make a dual axis chart over 4 categories of education level.

In the interests of examining job demand by geographical locations and business industries, I decided to make use of tree map and side–by–side bar chart. The conventional choice to visualise spatial data is through a map. In this case, a proportional symbol chart may seem to be a viable option, but I also want to evaluate job demand not only on country level but also on larger continent level and smaller area level. A dot on

a continent map would be expected to collapse into smaller dots representing countries within that continent when it is zoomed in on, and continuing to zoom in on a country should produce similar effects. However, apart from technical infeasibility, the number of countries in the dataset is only 16, each of which contains only a limited number of areas, so visualizing them all on the map would make it asthetically scattered. Tree map on the other hand helps achieve the task with the capability to drill down to lower levels while preserve the focus on a particular geographical level at a time. With respect to industries, the dataset contains 123 industries into which a company business can be classified. Similar to the data on skills, visualising 123 bars would cause the plot to be cluttered, but the ultimate goal is to identify top industries with highest demand, so sorting them in descending order is sufficient. However, in order to allow for close examination of industries in the middle or at the bottom of the chart, an area of bars can be zoomed in on by a brush.

Based on the concepts, types of data and statistical correlation, the visualisations can be grouped into 3 sections: job duties & skills, education & experience, and job demand. This requires tabset panel to navigate among these three sections. In addition, as going through the visualisations, users may want to have the results tailored to their preferences for a particular job level and position. Filtering is added to prompt user inputs, and all visualisations are expected to change accordingly. Specifically, there are two ways to process underlying data and display results upon requests to introduce additional dimensions. One way is to retain single visualisation while have the data aggregated over all selected categories, which is applied for word tree, tree map and side–by–side bars representing industries. The other way works for the remaining charts, which is to have multiple facets of the categories. These are the main interactive elements of the entire set of visualisation, which has consistent effect on all charts. However, there are also other interactive features that are applied locally to certain charts.

Implementation

Shiny is the backbone of the interactive visualization with its graphical elements supported by *ggplot, plotly* and *googleVis* along with independent packages *word cloud* and *gwordtree*. *googleVis* provides R interface to Google charts API that enables me to create JavaScript–based interactive charts, which in this project are Word tree and Tree map. The visualisation has 4 tab panels containing the information that addresses the previously described concerns of job seekers in the same order. The entire set of visualisations is controlled by the master inputs from users for experience level and job position, which is placed on top of all sections. This is mainly beacause these two attributes are highly associated with users' profiles and the majoriy of users are inclined to start searching for jobs with a particular position and level in mind. Users can select only one data position at a time but he can choose to view multiple experience levels and whether the displayed results will be aggregated and segregated depends on the charts. This mean that changing these input values in this section produces customisations to all of the following visualisations. Each section has their own interactive elements attached to certain figures that allow users to freely tailor the displayed results to their preferences. All of the inputs are pre–selected with a certain value.

The default tab serves two purposes: first is to allow users to eyeball the kinds of data I used in this analysis and second is to present details of the actual hiring positions to wrap up the job seeking process. This section contains no more than a list of job openings that match users' inputs for experience level and job positions. Another important job search criterion is geographical location, so country fitering feature is added to narrow down the results. The display of each data point resembles a typical job posting on LinkedIn, from which users can be directed to the detailed page of a particular job posting upon click on hyperlinks embedded in the titles.

The second tab contains information on job responsibilities and skill requirements. Word tree is used to highlight key activities of the selected role in which the first (parent) branches include verbs and the following branches refer to subordinate noun (phrases), which altogether best expresses the answers in the format "*To do what*". It basically shows the combinations of words or phrases that frequently go together and the size of the words implies the frequencies of each word in the dataset. To simplify data processing,

I made use of a supportive package *czxa/gwordtree* installed from GitHub to render the chart. The second half of this section illustrates Word clouds made up of keywords on technical and domain skills required in job descriptions. Like any other word clouds, the bigger a word, the more often a programming language, a software or a framework is required by employers. These two pieces of information are placed under the same section as they are both textual data and frequency is a useful metric to highlight the most important ideas.

The next section first provides an overview of the percentage of companies mentioning the number of years of experience in job descriptions with respect to each chosen job level. The remaining part aims to plot years of experience agains each categories of education level, factoring in experience levels. The two charts are both created by *ggplot*, and using *facet_grid* function for facetting over multiple job levels. The y–axis refers to the number of jobs at each degree for a particular job level, and while this is very straightforward for the bars, plotting another axis with respect to the lines for years of experience is challenging as they have different scales. This means that placing ticks on the second axis to the right requires rescaling the left axis as well, which would distort the trend of the line. Therefore, instead of dual axis, annotation is included above the line to provide the users with the true statistics. The trendline however alters with respect to the measures used to aggregate the data, so users are allowed to specify whether to aggregate it by mean or median. To create tooltips when users hover a bar or a point, I must convert *ggplot* object to a *plotly* object, which also offers more interactive options.

Switching to the last tab, users can explore the geographical locations and industries with highest job demand. To illustrate the variation among locations, a tree map visualisation is the most appropriate. This is due to the capability to drill down to view distributions at sub–region levels. Thanks to *googleVis* with built–in functionality to access deeper levels, tree map can be easily created with only a few lines of codes for data pre–processing and API calls. To compare demand across industries, the bar chart can be easily created with *ggplot* combined with interactive effects of *plotly*. Since there are more than 100 industries and it is hardly visible to plot all of the industries, top 5 industries are displayed by default and I leave it to the user to decide how many industries they want to show.

As far as the data is concerned, this visualisation is built upon 4 datasets, 3 of which have been pre–processed and aggregated to speed up the performance of the app. These are used for creating the word tree and word cloud diagram in section 2 as well as the bar chart in section 4. Meanwhile, the other charts derive data from the original dataset, and preprocessing are performed directly in the operation.

User guide

At the start of their exploration, users are required to have their devices connected to the Internet before being able to interact with the visualisation. It is also highly recommended that the app is activated in an open browser window as certain *googleVis* plot functionalities are only enabled upon external view. At the start of the visualisations presented 4 fun facts about the dataset displayed in the top circles. When hovering over a circle, users can view detailed explanation on the fact. Then users can begin to explore the visualisation by selecting any job levels and position of their interests.

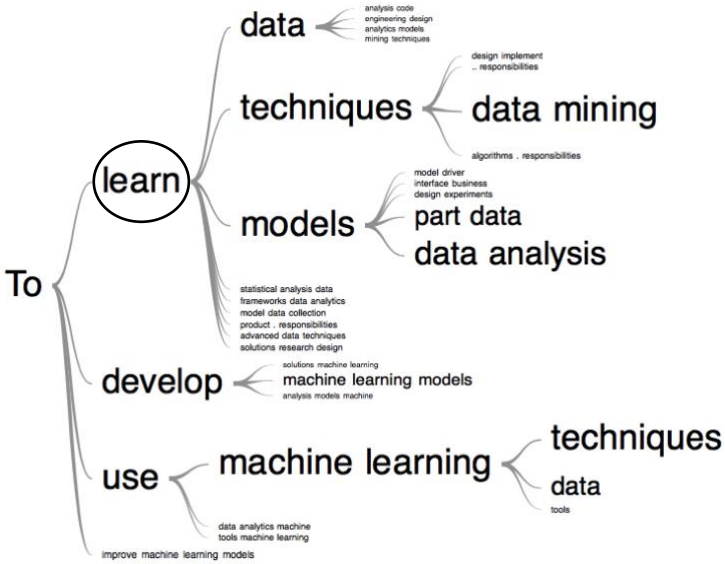
By default, the visualisation displays information on entry–level jobs for Data Scientist position, but users can always click the dropbox and a list of positions will be displayed available to choose from. While the options for data positions are single choice, users are allowed to select multiple experience levels at a time and how results will be displayed vary by charts. Below this filter section shown a summary of insights about the selected position from my data analysis, which aims to inform users of the types of information they can expect to find in this visualisation Located underneath are the 4 tab panels which can be accessed in no particular order. However, the current arrangement already follows the logic of job seekers that are explained above in the Implementation section. It is also useful note that altering the values for experience levels and job positions introduces changes to all the contents in each tab, or in other words the entire set of visualisations.

The first tab is selected by default, which presents the list of job postings in the dataset that match users' queries in the main filtering section. Apart from this, users can filter the results based on the countries where they wish to look for a job. Each posting is comprised of job title, organisation name, country/region and job type (Internship, Full time, Part time or Contract). Users are further allowed to view that posting in details by

clicking on the title in which embedded a hyperlink that will direct them to the LinkedIn job page. However, users must be aware that not all the links are still valid and many of the jobs are no longer available as the data was collected nearly 2 months ago.

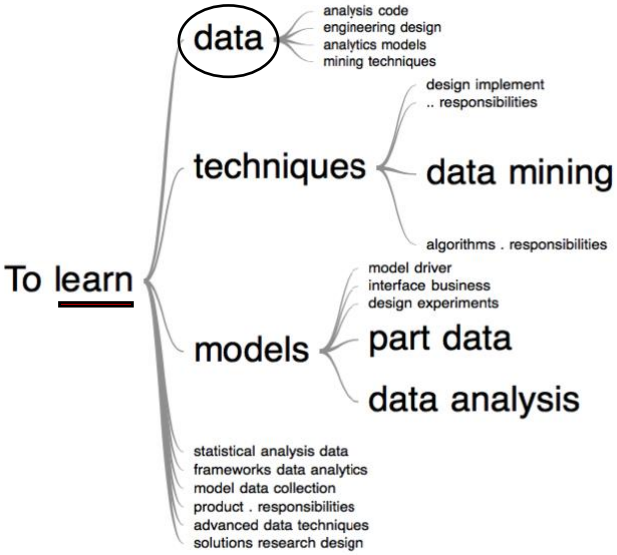
In the next tab, word tree diagram is used to highlight key responsibilities of the position while word cloud aims to show the skills more often required in a candidate including both technical skills – which refers to programming languages, tools or frameworks and domain skills – which are more or less correlated to the candidate's study area. The results shown in the word tree diagram is read from left to right. Each branch of the tree indicates an activity that an employee in that position is expected to do. In the following example, one of the main duties of an entry–level Data Scientist is to *develop machine learning models,* while another is to *learn data mining techniques.* To discover what else the candidate must *learn* during his/her job, users can simply click on the word *learn* and the diagram collapses on the sub–trees starting with that word. Users can continue to explore levels in deeper branches by clicking the parent nodes. As shown below, if users click the word *data*, the following branches from the phrase "*to learn data"* will be zoomed in at the center and they can understand more about the activity of *learning data* in this context. Returning to the root or parent nodes can be done by re–clicking the previously clicked word.

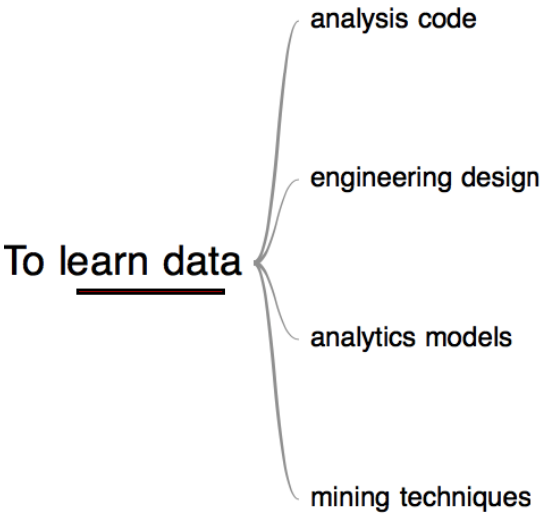
Data Scientist is expected



Click "learn" again

Click "learn"





Click "data"

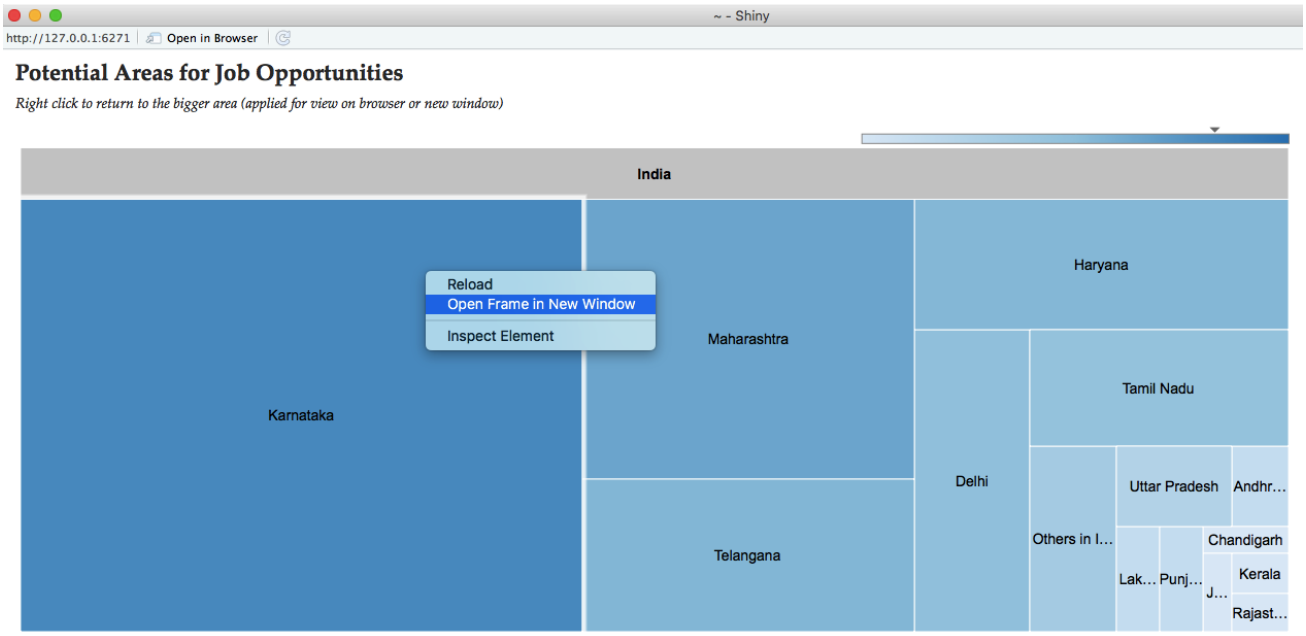
Click "data" again

In case multiple experience levels are selected, data will be aggregated and values will be updated to rebuild the tree. Also, in this visualization, hovering over a certain word shows its weight – the number of times it appears in the dataset, and the bigger a word or a certain phrase, the more likely it is an important role. Unlike word tree, distinct word clouds will be generated upon multiple selection for experience levels. The clouds are themselves static, but users can interact with them by adjusting the maximum number of words to be displayed. Word cloud is useful in providing users the instant answer to which skills they need to focus on developing. As in word tree, the size of words is proportional to their usage. The largest words around the cloud center refer to the most highly requested skills.

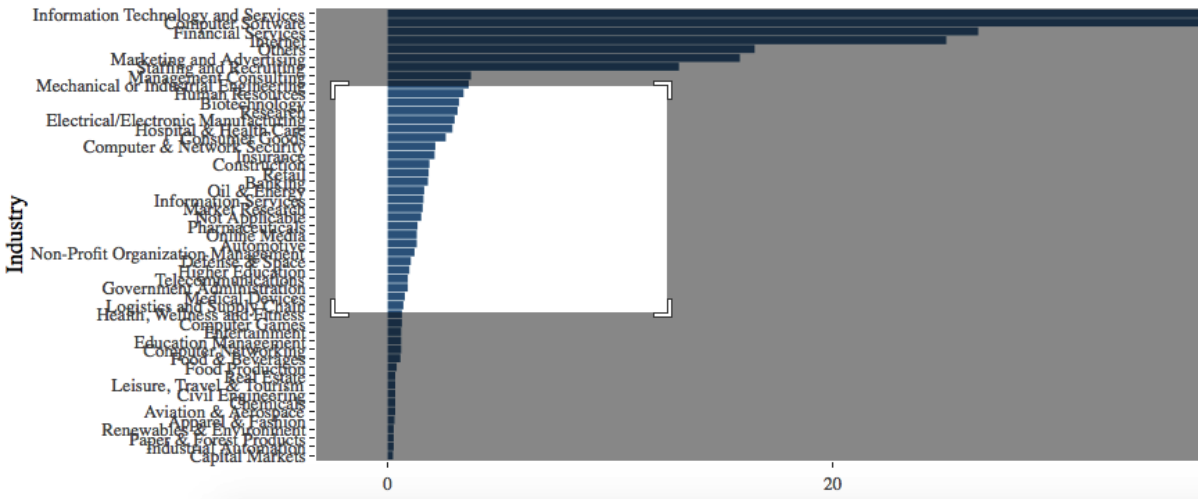
The next tab leads users to explore job requirements on education and experience. The first graph is self–evident with percentage values annotated directly on the bar. The combined bar–line chart below aims to illustrate the correlation between education level and years of experience. The bar represents number of jobs while the line describes the average years of experience with respect to each education level. A noteworthy statistical reminder is that the data used to illustrate years of experience only include companies that specify the number, so the samples of the bar and line chart are not the same. By default, years of experience are arithmetically averaged, but users are also presented with another option to aggregate it by median. This chart addresses such questions as how many companies requesting at least Master degree for an entry–level position or what is the minimum number of years of experience a qualified applicant should have. When multiple job levels are selected, data will be segregated and each level will have its own set of charts.

Moving on to the final tab, users will be presented with another visualisation of data tree – tree map in which each node is represented as a rectangle with sizes and colors corresponding to its value, and deeper levels can be accessed through drill down functionality. This means that users can move down the tree by left clicking a parent node and to move back up by right clicking any of the child nodes. At first, users can observe the distribution of jobs across continents in the world. When they click the rectangle of a continent, they will be able to compare job demand among countries and different regions within a country can be examined in a hinted fashion. However, this

right–clicking feature can only be activated if the app is viewed on a browser or a new window frame (*A new window frame can be initiated in R Studio by right–clicking the app and choose Open Frame in New Window as in the image below).*



Scrolling down to the bar plot, users can choose how many industries to be displayed in the same plot by draging the slider along the scale of between 20 (default value) and 123. By hovering over a bar, a tooltip appears showing the proportions of companies in a certain industry are hiring for the selected position. The proportions do not sum up to 100% as there are some companies operating in several industries. In cases users want to explore industries with low demand, they can highlight a selected area using a brush to zoom in on corresponding bars (as shown in the image below). Users can simply go back by clicking anywhere in the graph.

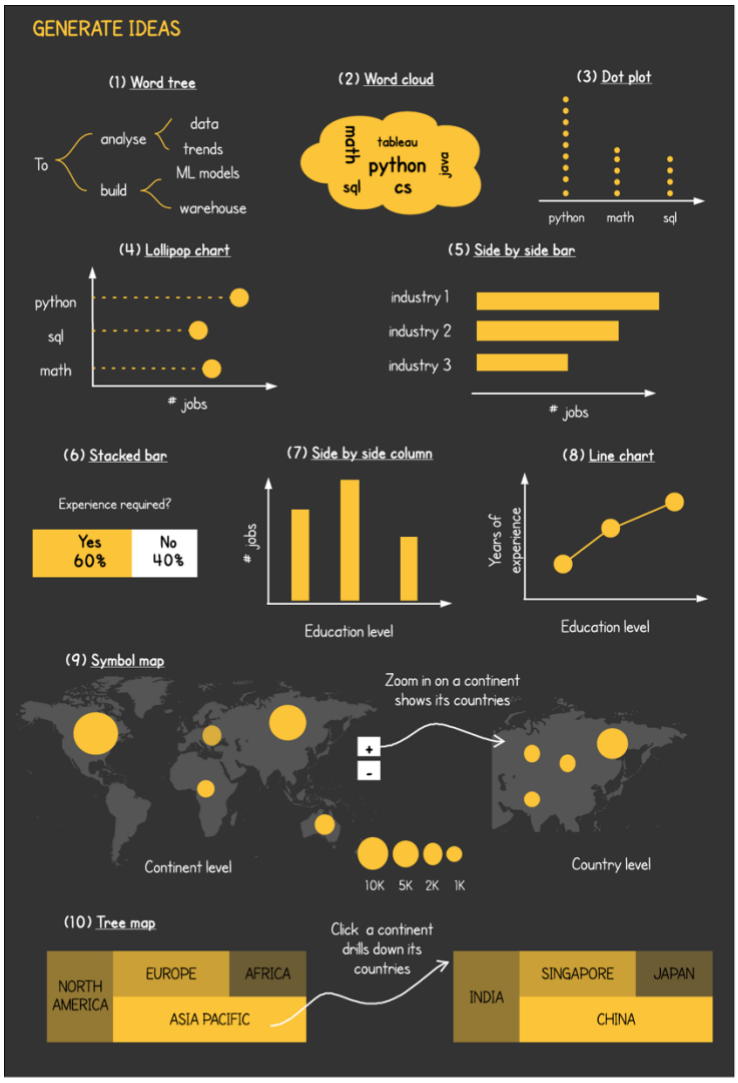


Conclusion

The visualisation aims to provide sufficient and useful insights to assist future Data Scientists in seeking potential jobs and making sound career decisions through a range of self–explantory charts with experience–enhancing interactive elements. Though achieving the tasks, the visualisation still has some disvantages. Word tree is prone to errors during data pre–processing which involves tagging parts of speech to extract sequences of words that must at least contain a verb and a noun. Handling such textual data is the most challeging task in this project, which must be done with extreme care as contexts are severely lost when phrases are extracted from the large document of job description, which strongly affects the quality of displayed information. Though highly effective when plotting related keywords, available word cloud package in R fails to allow users to examine low–frequency words – for example – by selecting maxium frequency or brush options. The fact that the entire data is textual limits the choices of visualisations. The only numeric data is years of experience, which is however subject to large number of null values. This greatly reduces the size of sample used to calculate the average number of years, which are later represented by a line chart. The figures may be not reliable, thereby distorting the trend. The learning point is that it is essential to assure the quality of data prior to visualisation. Otherwise, it would be meaningless and sometimes convey misleading information. With regard to implementation, the libraries are extremely convenient thanks to their built–in interactive elements, but at the same time this causes the design to be restricted and inflexible. Given the time and technical competence, I would have opted for D3 and JavaSript or incorporated more D3 functionalities into R, which would have added more flexibility and interactivity in the design.

Appendix

Sheet 1



Visualised concepts

1. Job duties
2. Job requirements
   a. Skills b. Education c. Experience
3. Job demand
   a. By countries b. By industries

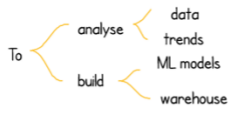
CATEGORIZE

Concept 1 + Concept 2a	Concept 2b+ Concept 2c	Concept 3
1, 2	6, 7, 8	5, 10

COMBINE & REFINE

Tab 1: **Job duties & Skills**

(1) Word tree

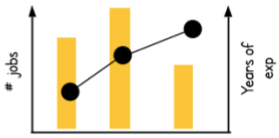


(2) Word cloud

Tab 2: **Education & Experience**

(6) Stacked bar

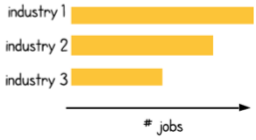
(7+8) Combine bar & line



Tab 3: **Job demand**

(10) Tree map

(5) Side by side bar



FILTER

Keep: (1) (2) (5) (6) (7) (8) (10)

Remove: (3) + (4)

Duplicated contents with chart (2) but not as efficient because there are more than 100 skills, which would be cluttered if all are plotted.

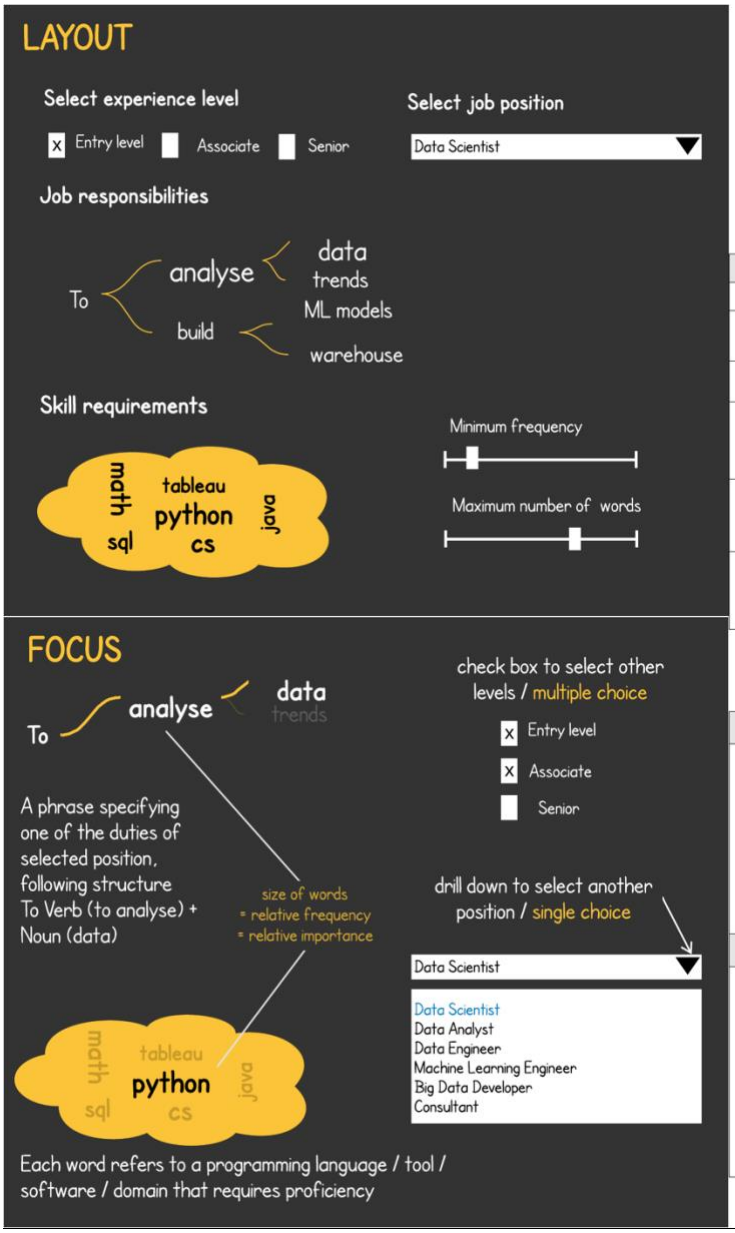
Remove: (9)

Dataset contains only 16 countries, which contain a limited number of regions. The map would be sparse

QUESTIONS

How can we visualise the concepts over different experience levels and job positions?

Sheet 2



Title: Data Science Career Insights
Author: Tran Vo
Date: 3 June 2020
Sheet: 2 - Initial design 1
Task: Visualise tab 1 - Job duties & skills

OPERATION

Action	Word tree	Word cloud
Select a position	Values updated	
Select multiple experience levels	Values aggregated. Only 1 tree displayed	Faeting over multiple levels
Hover over a word	Tool tip displayed showing frequencies of the word	
Left click a word	Branches rooting from that word zoomed in. Click again to go back	No effect
Slide to value *k* on "minimum frequency" scale	No effect	Only keywords appearing at least *k* times displayed
Slide to value *k* on "maximum number of words" scale	No effect	Only *k* most frequent keywords displayed

DISCUSSION

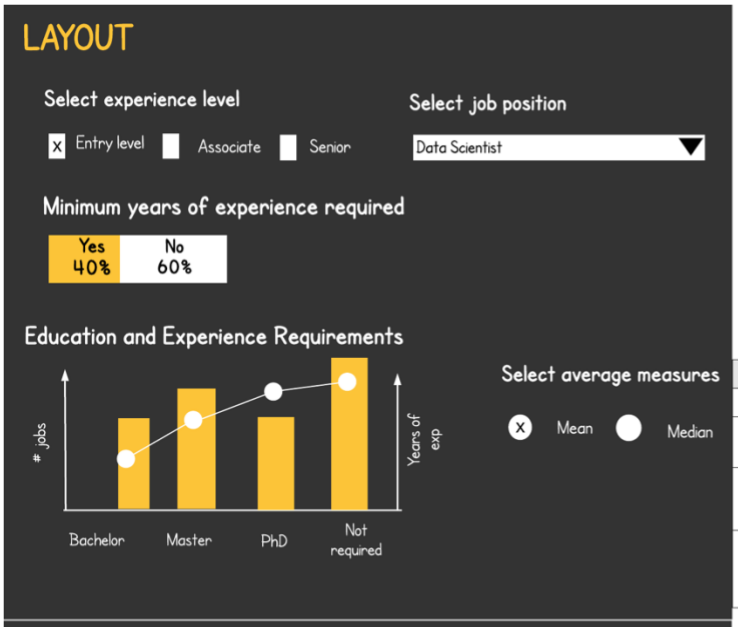
Word tree

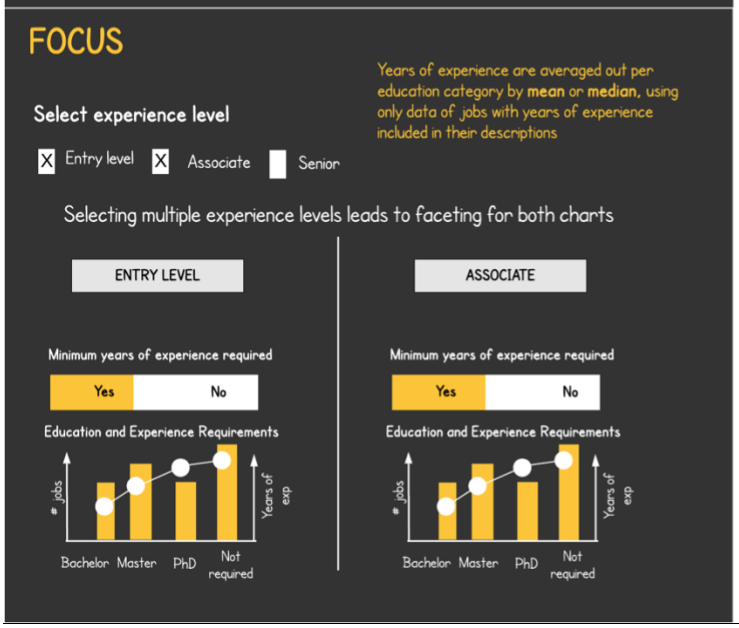
Impossible to compare job responsibilities different over experience levels
Branches in word tree with a large number of child nodes (k) tend to collapse to tendrills labelled "k more"
Text cleaning has a large effect on the displayed results
Phrases are more informative than simple keywords

Word cloud

Possible to plot all 130 skills at once
Straightforward intepretation
Heavy data wrangling to extract skill-related keywords
List of keywords is not exhaustive
Impossible to select maximum frequency through R wordcloud package -> difficult to examine less frequent keywords

Sheet 3





Title: Data Science Career Insights
Author: Tran Vo
Date: 3 June 2020
Sheet: 3 - Initial design 2
Task: Visualise tab 2 - Education & Experience

OPERATION

Action	Stacked bar	Combined bar & line
Select a position	Values updated	
Select multiple experience levels	Faceting over multiple levels	
Hover over a bar / point	No effect	Tool tip displayed showing data values
Select a average measure	No effect	Line values updated, showing years of experience average by chosen measure (mean or median)

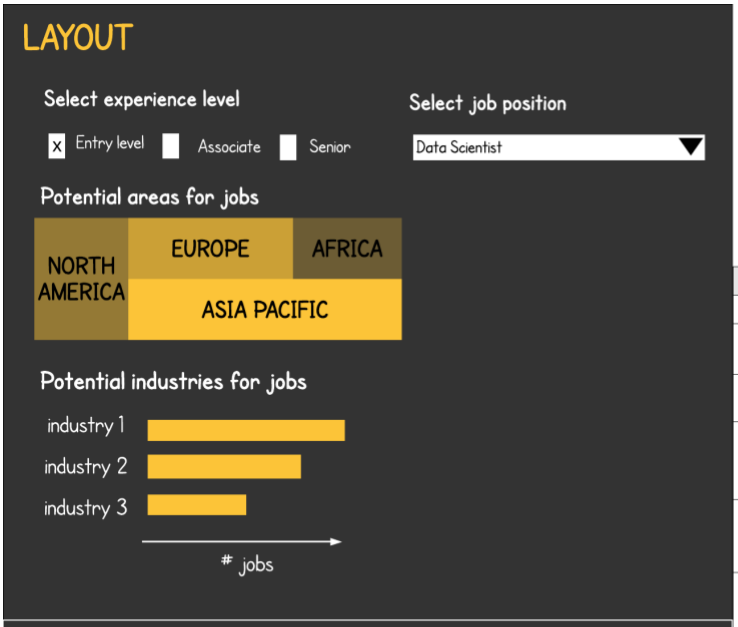
DISCUSSION

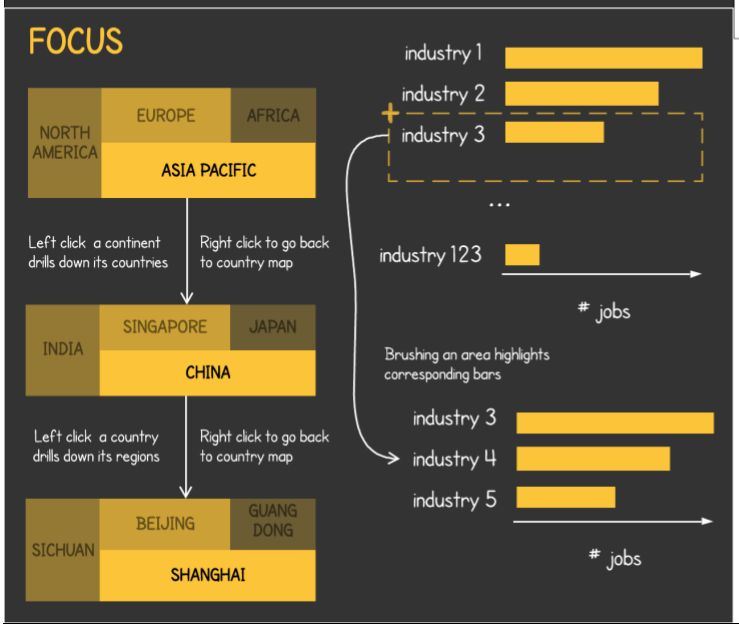
Possible to investigate relationships among 3 variables: experience level, experience and education level

There are far more companies not specfiying years of experience in job descriptions than those that do -> average values may be not representative and line patterns would be misleading

Plotting the right axis with different scales by ggplot requires rescaling the original left axis -> difficult to find the right "ratio" to balance the size of bars and reletive positions of points.

Sheet 4





Title: Data Science Career Insights
Author: Tran Vo
Date: 3 June 2020
Sheet: 4 - Initial design 3
Task: Visualise tab 3 - Job demand

OPERATION

Action	Tree map	Side by side bar
Select a position	Values updated	
Select multiple experience levels	Values aggregated. Only 1 chart displayed	
Hover over a bar/node	No effect	Tool tip displayed showing data values
Left click a bar/node	Move down the tree	No effect
Right click a bar/node	Move back up the tree	No effect
Brushing over an area	No effect	Zoom in on bars within the brushed area

DISCUSSION

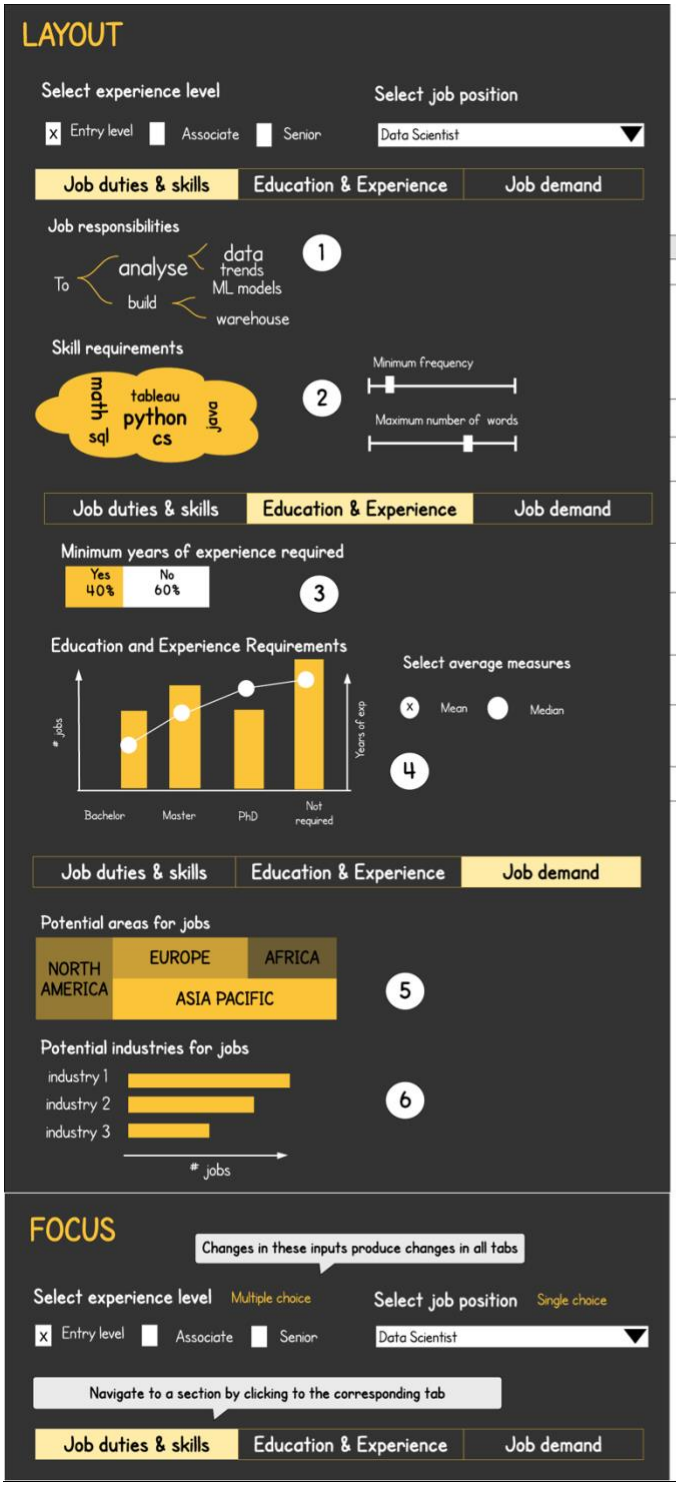
Tree map features alllow for comparisons among geographical locations on 3 different levels

Brushing features allow for examination of less popular industries

Plotting 123 industries by default is not aesthetically desirable

Impossible to compare job demand by each experience level

Sheet 5



Title: Data Science Career Insights
Author: Tran Vo
Date: 3 June 2020
Sheet: 5 - Realisation
Task: Complete layout

OPERATION

Action	Charts applied	Effects
Click a tab	All charts	Navigate to visualisations in chosen tab
Select multiple experience levels	All charts	1 chart displayed with values aggregated (1) + (5) + (6) Faceting over multiple levels (2) + (3) + (4)
Select a position	All charts	Same chart(s) displayed with values updated
Hover over a bar/node/word/point	(1), (4), (6)	Tool tip displayed showing data values
Left click a bar/node/word/point	(1), (5)	Move down to sub trees (1) + (5) Move back up the tree when it's clicked again (1)
Right click a bar/node/word/point	(5)	Move back up the tree
Slide to value *k* on "minimum frequency" scale	(2)	Only keywords appearing at least *k* times displayed
Slide to value *k* on "maximum number of words" scale	(2)	Only *k* most frequent keywords displayed
Select an average measure	(3)	Same line chart(s) with years of experience values averaged out by chosen measure (mean or median)
Brushing over an area	(6)	Zoom in on bars within the brushed area

DETAIL

Datasets:

- Data on action phrases (Verb + Noun) extracted from job descriptions
- Data on skill-related keywords extracted from job descriptions and total frequencies for each keyword (including technical & domain skills)
- Data on business industries and total number of jobs for each industry
- Data with additional attributes on education level, years of experience and geographical areas (including continents, countries, states/areas)
- Segregate each dataset by Experience level and Job position

Dependencies: R, Shiny, googleVis, ggplot, ggplotly

Requirements:

- Connected to Internet
- View & interact with visualisation on browser / open window