MSIN0167

Data Visualisation

Group Coursework

Team

OCULART

Word Count: 3997

11TH June 2020

Introduction

Data visualisation is gaining more value nowadays. It is estimated that the total market value for data visualisation can reach up to 8 billion dollars by 2023 (MRW, 2018). Kirk (2016) defines data visualisation as "the representation and presentation of data to facilitate understanding". Although many argue that visualisation is subjective to a large extent, Kelleher and Wagener (2011) also stress the importance of guidelines and frameworks on conveying information effectively through visualisation.

In this report, we aim to present two visualisation presentations (a double-A4 page static and an html interactive visualisation) of a chosen dataset based on a 4-stage design process (See Figure 1).

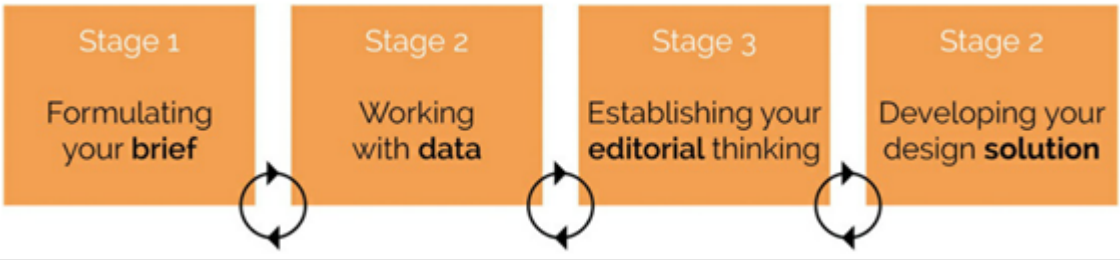


Figure 1: The Four Stages of the Visualisation Workflow (Kirk, 2016)

Step 1: Formulating your brief

1.1 Curiosity

Curiosity should be the starting point of a data visualisation project (Kirk, 2016). Curiosity can help us determine what is the most valuable and relevant analysis when we are working on a dataset.

As MSc students studying Business Analytics, many of us consider data-related roles (Data Analyst, Data Scientist, Data Engineer) as a future career. Although, the size of teams in these disciplines may differ across companies of the same sector or even of the same size.

Companies which recognize the value of data-driven operations may have already established data teams, and they may invest a lot on them. On the other hand, there are also companies which started recently to recognize the value of their data, and they still have relatively small data teams compared to the size of their operations.

Hence, our curiosity focused on the unique characteristics that describe small and large data teams.

1.2 Circumstances

Circumstances refer to all the requirements and restrictions that are inherited by us, imposed on us or determined by us (Kirk 2016). These circumstances define our project. An extended analysis of these parameters can contribute to well-informed decisions for the subsequent

steps. The circumstances may comprise of five aspects: people, constraints, consumption, deliverables and resources. Below we analyse each of them.

1.2.1 People

This aspect refers to all the affiliates with our project. These affiliates may include the intended audience as well as (content) creators of our team.

- Audience

  The audience for our project it can be people who work in data-related positions or people who work in IT industry and they would like to change move their career in a data-driven position. Besides, HR managers may want to have access to our presentation. The findings can help them to better define job vacancies (proper job titles), understand the skillset of different roles.

- Creators

  For our project, the actual creators are the three members of our team. We expect that our existing knowledge in the data-related fields but also the theoretical and technical knowledge that we have gained through the data visualisation module can contribute to the successful completion of the project.

1.2.2 Constrains

Any project may have observable or unobservable constraints which may limit our ideas our final visions for a project. Below we analyse some of the potential constraints that we may face in data visualisation projects, and we extend these in our project.

- Timescales

We believe that project management tools (JIRA, Microsoft Teams) have contributed to better managing our available time (as they can better allocate tasks to the team members and reduce idle times).

In our case, the start date of the project was 20th of May 2020, and the end date is 10th of June 2020. The milestones that we have set-up with the team can be found in Appendix 1.

- Pressures

In our project, we believe that we do not face any direct pressure from our internal or our external environment.

- Design

For our final deliverable, there might be design restrictions. Depending on the medium that will deliver our project (newspaper, web, banners), we may need to use different colour schemes or adjust the size of the visualisations to be easy to read. Also based on the audience that we intent our project for, we may need to use a different design (for

example different graphics may be used for a professional document compared to a document which intends to inform a group of students).

- Technological

For our project, we recognize that there might be other software tools that we are not able to test at the moment (this also refers to timescale constraints). Regarding hardware, we believe that we will not face any limitation, as our sources (data files) are relatively small compared to our available resources (computation power).

1.2.3 Deliverables

Regarding the way that the deliverables are consumed we find that frequency and setting play a pivotal role:

- Frequency

For our project, we do not expect that the deliverable will be replicated across future periods. Although significance will be given to the documentation of our solution, so it will be easy for anyone who is interested in replicating our solution.

- Setting

For our case, we do not expect that the material will be delivered with any additional help. For this reason, we aim to offer a solution that can be conceived without the need of any other intermediate.

In addition, we do not have to pass our messages quickly to the audience. We may consider people who do not wish to spend much time on reading our analysis (probably with highlighted statistics), but in general, we expect that they can have a look on our analysis at their own pace.

Regarding the nature of the deliverable, we need to consider the medium and the quantity of it.

- Medium

In our project, we will deliver a static presentation for a magazine and an interactive visualisation for a webpage. For the first we assume that the magazine is printed in a high-quality glossy paper, so the colour rendering will be unlikely to be affected. For the interactive visualisation, we assume that the magazine will be presented properly on web-browsers of at least 1280x720 pixels (720p). Ideally, we would like our visualisation to be of high resolution so it can satisfy presentations also on screens with high resolutions or points per inch (ppi).

- Quantity

It refers to the number of elements for each deliverable. In other words, how many visualisations will be included in the deliverable. This is a question which cannot answer in with certainty in advance, but we can have a rough estimate for it. It is worth to mention

that quantity it seems to be correlated with design constrain of our analysis. For example, for two A4 pages, we may produce 4 to 6 visualisations. However, for a higher number of visualisations we may violate other constrains, or we will not be able to fulfil the scope of our project.

For our static deliverable, we produced 6 visualisations (one is a static version of an interactive one), and for our interactive deliverable, we produced 5 more visualisations.

1.3 Purpose

Kirk (2016) points out that the definition of the purpose of a visualisation has a significant impact on the viewer's understanding process. Hence, it is necessary to carefully design the purpose of the visualisation.

We believe that having a better understanding of our future career should be extraordinarily prudent. Therefore, we would like to convey critical information to broaden people's knowledge of data teams of different size. It ultimately could help an individual to realize where could be a better fit.

We also acknowledge that apart from people who would like to join small or large data teams, our visualisation project can also provide valuable insights for different stakeholders such as employers and HR managers. To sum up, the purpose of this visualisation project is to provide stakeholders with valuable insights regarding the different size of data teams.

In terms of the experience of the visualisation, we decided to use “Explanatory” experience. For the static graph, we are trying to bring critical insights to the surface to facilitate understanding as the page size is limited (I.e. maximum two pages) and no interactive features can be used on printed magazines or newspapers. As the web version of the visualisation serves as an extension of the static visualisation, we chose to continue use “Explanatory” experience and provide a few functions to let viewers find some information that they are interested in.

In terms of tone of voice, we decided to use the reading tone. There are two main reasons why we believe the reading tone would be appropriate. Firstly, the reading tone can help viewers perceive data efficiently, especially in the static visualisation, where the page size is limited. Secondly, the reading tone is in line with our purpose, which provides accurate and insightful information for stakeholders.

1.4 Ideas

Before we conduct the data exploration, we need to define principles for creating our visualisations. We believe that we need to demonstrate insights that are more general at first; however, they must be relevant to our audience's needs. Hence, the first few plots should contain multiple colours to attract viewers' attention, and lengthy text annotations should be avoided. Then, we should gradually focus on more detailed visualisations and provide annotations that provide extra information when viewers are interested in the visualisations and are willing to absorb more information.

For producing our visualisations, we got inspired by popular magazines such as Financial Times and the Economist. We aimed our visualisations to follow similar aesthetics but also annotations which can make our audience to focus on specific insights.

Step 2: Working with data

2.1 Acquisition

The dataset we have chosen is the "Kaggle's State of Machine Learning and Data Science 2019 survey". This annual survey collects a wide range of information of users who work with data. It includes questions which cover educational background, employment and professional tools used in the workplace. Our visualisation solution will explore what the different characteristics of small and large data teams are.

It is worth mentioning that we had a few potential datasets to choose from before our final selection. At first, there were four potential datasets to choose from. For all of them, we noted their strengths and weaknesses. In order to identify the dataset that had the most potential, we created a selection matrix (See Appendix 2) which contains strengths and weaknesses of each dataset, followed by a possible editorial theme and 3 selection criteria (Technical Requirement, Story-telling Requirement and Audience Profiling Requirement).

Among the three selection criteria, the Technical Requirements is the most critical one. We believe that the dataset should include an adequate number of attributes (i.e. more than 30), thus allowing us to create different graphs and focus on different aspects. Ideally, it should include geolocation data (geographic areas) & repeated instances (rows) across different periods for the same objects. Timestamps should also be included in an ideal situation. This would allow us to show not only comparisons but also create appealing stories. (e.g. how x behaves across different countries (map) / how x behave across years/months/days). Finally, the dataset should include different categorical variables (i.e. more than 3).

After summarizing strengths and weaknesses, we found out that the Kaggle ML & DS is the most suitable choice based on our criteria.

Although we decided to provide visualisation solutions for a data-related topic, we acknowledge that there might be better datasets regarding our topic. Hence, we searched related datasets on Kaggle and sorted them by relevance. We carefully examined other relevant datasets (mainly previous surveys), and we found out that the Kaggle ML & DS Survey 2019 still outperforms other alternative options. The main comparisons were among the size of datasets, number of attributes and the number of instances.

2.2 Examination

The Kaggle's State of Data Science and Machine Learning 2019 dataset that we have used for our project consists of responses from 19,717 Kaggle members from 171 countries.

The actual number of survey questions is 34, but for some questions, the answer is divided into several columns, so the total number of columns is 246. The responses to each question are composed of numeric data, categorical data, and text data.

2.3 Transforming

After loading the dataset, we firstly assessed all questions. It is an essential step as the original dataset contains 246 columns derived from 34 questions. We then loaded datasets from previous years (2017 and 2018) and did the same step to find any matching questions. However, it turns out that the format of questions varied across years, which means that it was difficult to join the questions from the past three years together and analyse the trends.

As the dataset contains geographic information (e.g. country of origin), we then performed data transformation on country names to make them more interpretable but also match them with the ISO format that Plotly package uses. It is worth mentioning that we excluded data regarding respondents who do not work in teams (e.g. team size = 0 or 1) when we generated our crucial attribute.

2.4 Exploration

Our goal is to deliver a comprehensive insight regarding data experts. Since the dataset we selected contains multiple questions and responses, we contemplated that it is necessary to select a key attribute and use it as a foundation of our project.

After several attempts, we considered team size as the basis for our project. We categorized respondents into two groups: small groups (1-4 people) and large groups (> 4 people), and the data are almost evenly distributed for the two groups, so it was suitable to be used as a critical attribute. After setting the team size as the key attribute, we then explored and compared the variations of other attributes between the two group sizes (for example salary gap in small teams and large teams).

Step 3: Editorial thinking & Developing our design solution

3.1.1 Common Editorial Theme: Employed in Small and Big Data Teams

Common Angle: What are the different or common characteristics between the employees in small and big data teams.

Common Framing: Only the respondents from the survey who currently work (11743). Within this group, we excluded the respondents who work alone, or they are students. In addition, we excluded those who did not give an answer to the size of their data team.

Common Focus: On two distinct group that they were selected after some trials; The first group referred to as "Employees who work in small teams" regards employees in a team of at most four members. The second group referred to as "Employees who work in large teams" regards employees in a team with more than four members.

3.1.2 Developing our design solution

The detailed breakdown of our proposed design choices will be covered right after the editorial thinking explanation of each plot (starting from section 3.2). It is worth mentioning that for static plots, the "Interactivity" part will not be covered as there are no interactive functions in the static presentation.

Common Composition: our guiding principle is that we would like our viewers to see the annotations in the charts clearly when it comes to the optimum size of plots. Hence, before we started to visualize our results, we set the figure size to be (8,8) with figure dpi at 300 to ensure a high printing quality of the figures. It is also worth mentioning that we will enlarge or change the aspect for some of the plots proportionally when we try to fit them in final presentation output to achieve an appealing visual presentation.

Common Colour: Inspired by the visually appealing plots from Financial Times, we applied "Solarized Light" (RGB: (255, 244, 220)) to set the base layer of all our plots. Furthermore, we used Adobe Color to generate the colour palette of the visualisations.

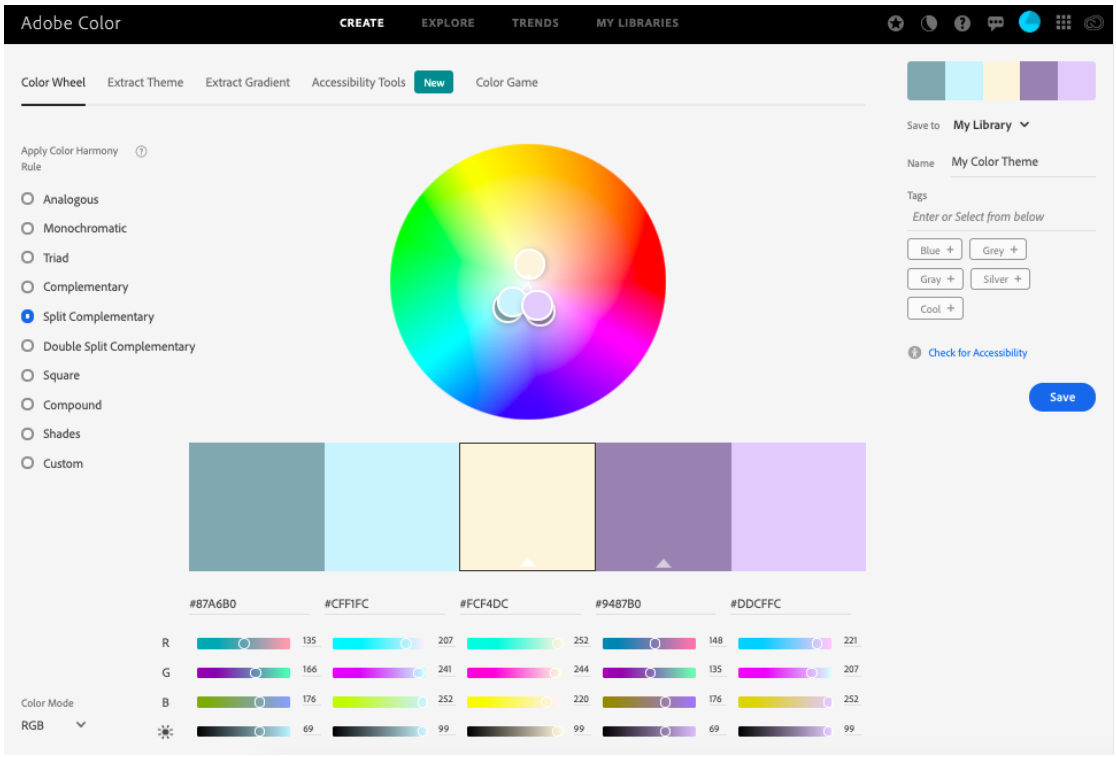


Figure 2: Screenshot of Adobe Color

3.1.3 Deliverables

Static Presentation

We used an online magazine editor called Canva to create a magazine-style static presentation which includes all the static visualizations above. We also provided description of each static visualization in the presentation. The presentation can be found in the high quality pdf (300dpi) of this document's directory or in the appendix 2 of this document.

Interactive Presentation

The interactive presentation was developed with Matplotlib, Seaborn & Plotly libraries and D3.js . You can access the final presentation on the website oculart.surge.sh. The source files can be found on glitch glitch.com/~oculart

3.2 Detailed explanation of the graphs we have chosen

3.2.1 Static Visualisation 1

Angle: What is the allocation of employees of different educational background across small and large teams?

Framing: From our initial analysis, we found that the respondents could select one of the following answers:

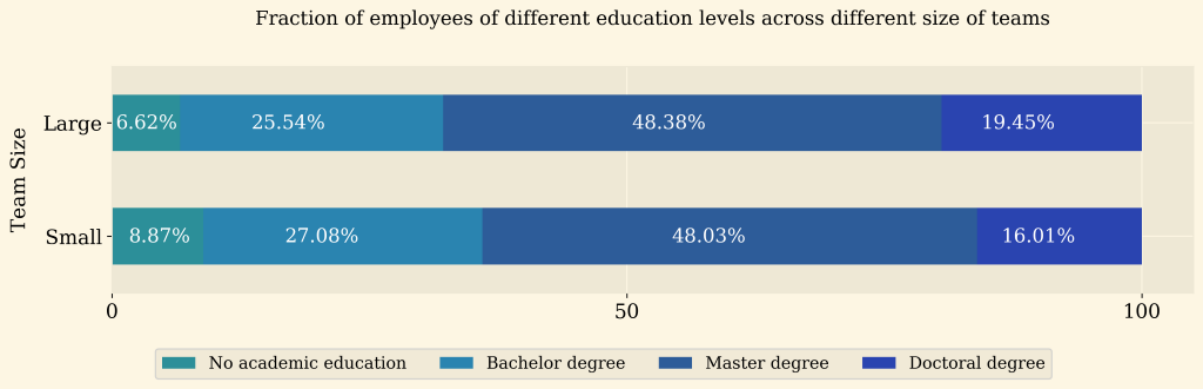
Possible answer	Total Responders
Master's degree	5586
Bachelor's degree	3039
Doctoral degree	2073
Professional degree	432
Some college/university study without earning a bachelor's degree	350
I prefer not to answer	160
No formal education past high school	103
Total	**11743**

From these respondents, we removed those who preferred not to answer.

Focus: For making our visualisation easier for interpretation, we decided to create 4 general groups with these categories (No academic education, bachelor's degree, Master and Doctoral degree). The category "No academic education" keeps all the respondents with the answers "Professional degree", "Some college/university study without earning a bachelor's degree", "No formal education past high school", which equals to 885 respondents. Please note that we have removed the respondents who preferred not to answer (160 respondents).

Before finalizing our translation, we grouped the respondents by the size of their team (small, large). Besides, for making a stacked bar chart, we translated the actual number of respondents to a fraction (in terms of the size team).

Annotation: We put the percentage values into each education level cell to help our viewers better understand the education level distribution across the two sizes of teams.



3.2.2 Static Visualisation 2

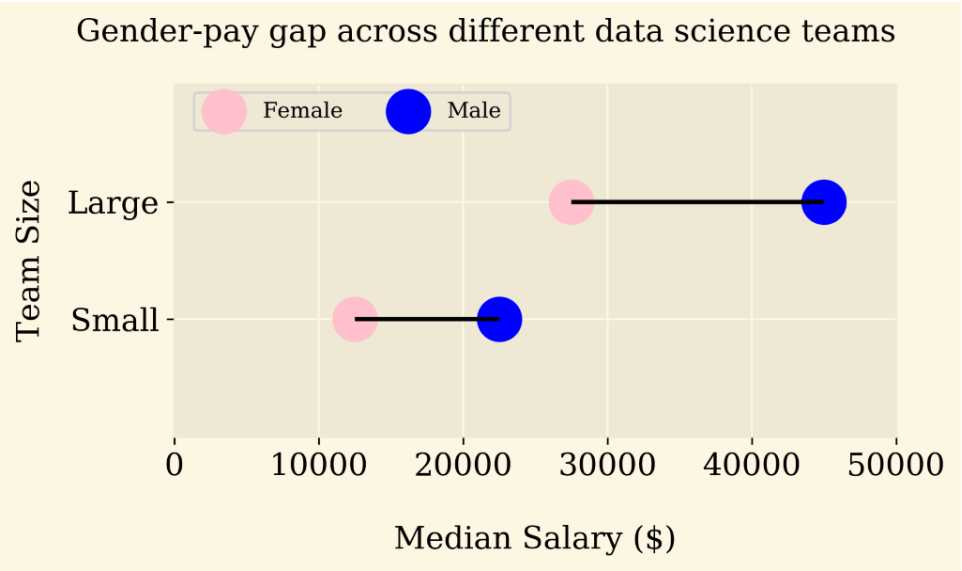
Angle: What is the median salary pay gap between genders across small and large teams?

Framing: First, we excluded the respondents who answered "Other" in terms of gender. The proportion of this group of respondents is so small that we can ignore it. Then we grouped respondents based on their gender and the size of the team that they belong to. For each subgroup, we calculated the median salary in USD. Below you can find the results:

Number of respondents	**Female**	**Male**
Big	27500	45000
Small	12500	22500
Total	40000	67500

Focus: We intended to not only find the median salary gap between genders within one team, but we also want to explore how big the salary gap is between two groups (small vs large).

Annotation: We provide a legend of the plot to help our viewers to be able to distinguish between the two genders.



3.2.3 Static Visualisation 3

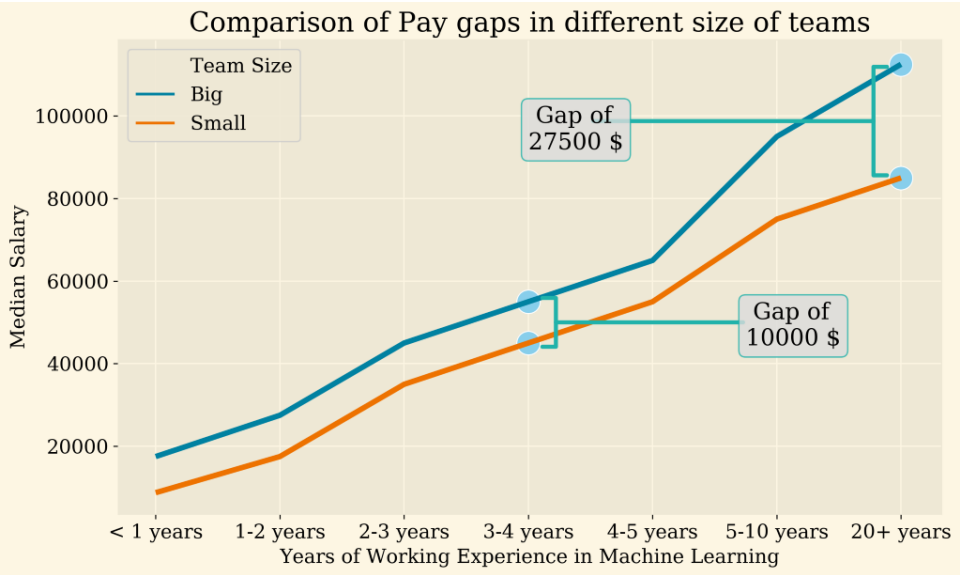
Angle: What is the impact of years of working experience in machine learning on median salary across the different size of teams?

Framing: We grouped the respondents into 14 groups in terms of years of working experience in machine learning but also if the team that they work is small or large. For each group, we calculated their median salary.

	Team Size	Years of Working Experience in Machine Learning	Median Salary
0	Large	< 1 years	17500
1	Small	< 1 years	8750
2	Large	1-2 years	27500
3	Small	1-2 years	17500
4	Large	2-3 years	45000
5	Small	2-3 years	35000
6	Large	3-4 years	55000
7	Small	3-4 years	45000
8	Large	4-5 years	65000
9	Small	4-5 years	55000
10	Large	5-10 years	95000
11	Small	5-10 years	75000
12	Large	20+ years	112500
13	Small	20+ years	85000

Focus: Instead of only displaying the median salary across different years of experience, we also annotated the pay-gap for employees with 3-4 years of experience with those of more 20 years of experience.

Annotation: We created two annotations to help our audience better understand the pay gap when respondents have different years of working experience across the two sizes of teams.



3.2.4 Static Visualisation 4

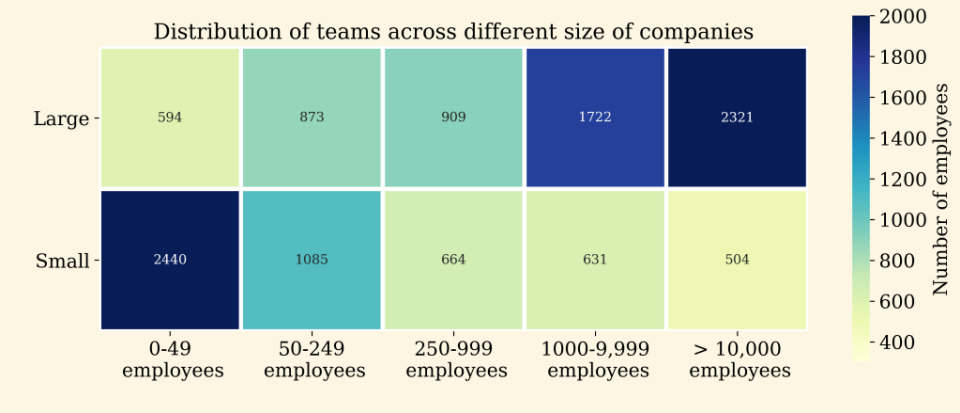
Angle: How the size of the company is related to the size of a data team?

Framing: We categorized the company size into five distinctive groups: 0-49 employees, 50-249 employees, 250-999 employees, 1000-9999 employees and 1000-9999 employees. Then we

	0-49 employees	50-249 employees	250-999 employees	1000-9,999 employees	> 10,000 employees
Large	594	873	909	1722	2321
Small	2440	1085	664	631	504

Focus: We intended to find which specific size of the company generate the most small/big teams, more importantly, we would like to use different colours to clearly show the position of the company size and enable our viewers to perceive it when they see it at first glance.

Annotation: Apart from a colour that helps viewers identify which values are extremes, we also provided the corresponding numbers of each heatmap cell to help viewers be able to quantify the impact of company size on the forming of different size of teams.



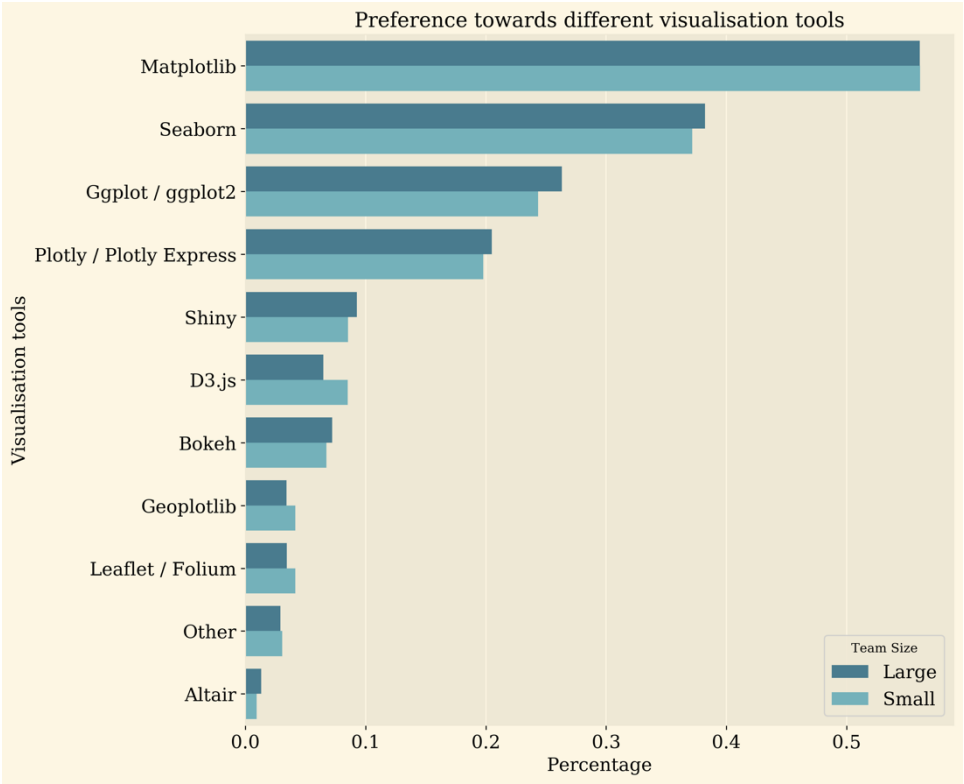
3.2.5 Static Visualisation 5

Angle: Which data visualisation tool is mostly used in the different size of groups?

Framing: We used the "groupby" function to generate a data frame that shows the absolute number of each visualisation tool used in the two sizes of teams. Moreover, we excluded the respondents who answered "None" in this question as this piece of information is irrelevant to our angle.

Focus: We would like to show which visualisation tool is preferable in both teams; we also intend to show the comparison of the visualisation tool usage between two teams. As the two sizes of teams have a different number of respondents, we transformed the absolute number into a percentage to make it feasible to compare the usage of one specific type of data visualisation tool between two teams.

Annotation: We provide a legend of the plot to help viewers to be able to distinguish between the two bars.



3.2.6 Static Visualisation 6 (Interactive Visualisation 1)

NB. This radar chart will be used in the interactive presentation as well.

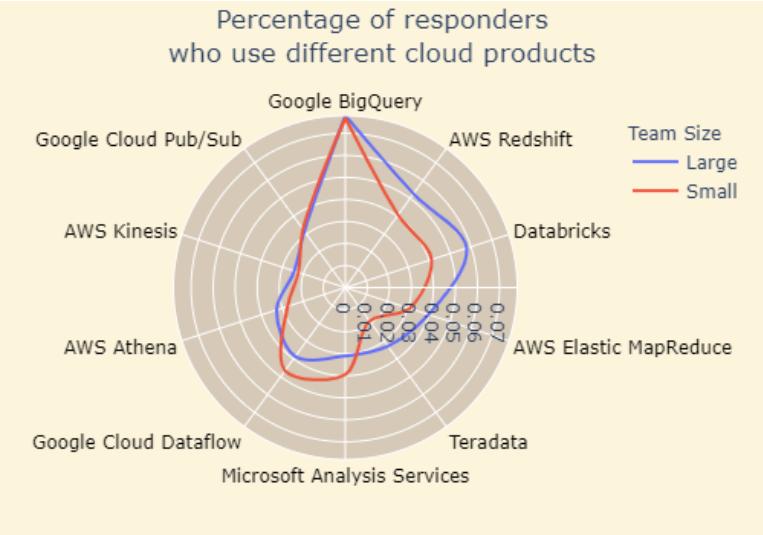
Angle: Which big data and analytics tools are more popular between the two sizes of teams?

Framing: We generated a data frame that contains the number of respondents under each choice across the two sizes of teams. Furthermore, we removed respondents who answered "None" and "Other" as this piece of information is irrelevant.

Focus: We transformed the absolute number into a percentage to make the graph more understandable. Also, as we already used some bar plots in previous visualisations, we intended to visualize our results by using a different type of graph.

Annotation: As this radar chart is easy to interpret based on the hit colours and percentage label, we did not provide any annotation for this chart.

Interactivity: Viewers can focus on one specific group by double-clicking the legend.



We used an online magazine editor called Canva to create a magazine-style static presentation which includes all the static visualizations above. We also provided description of each static visualization in the presentation[1].

[1] The static presentation can be found in appendix.

3.2.7 Interactive Visualisation 2

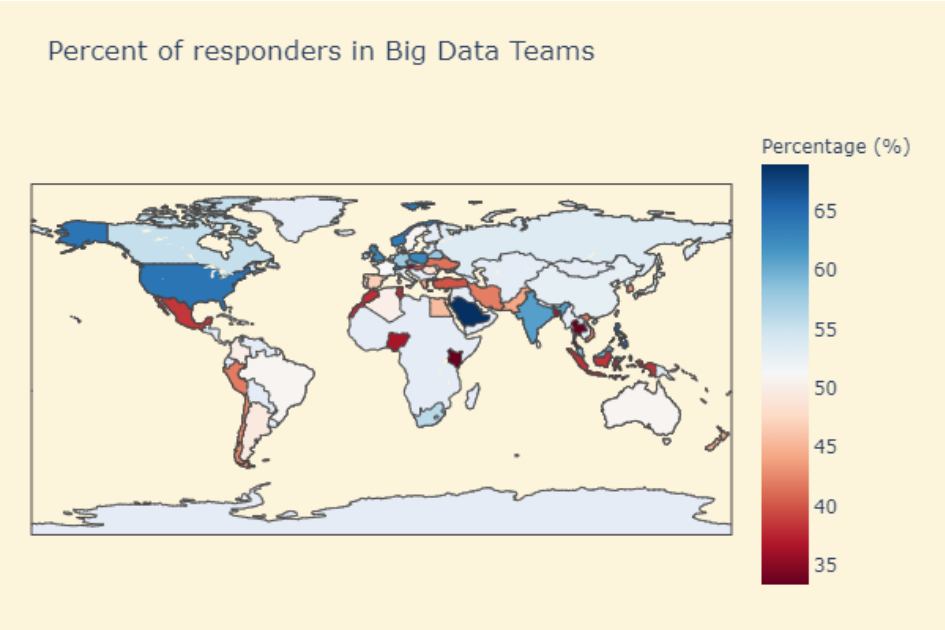
Angle: What is the percentage of respondents in large teams across different countries?

Framing: We used “groupby” function to generate a data frame that contains the percentage of the respondents’ nationalities in two size of teams.

Focus: We intended to find the connection between the distribution of large teams and countries. As large teams and small teams sum up to 100%, readers can easily translate the information to their own curiosity.

Annotation: We provided a colour bar so that viewers can use the difference in colours to identify which countries have a higher percentage of respondents in large teams.

Interactivity: Viewers can zoom in and hover on any country to find the percentage of respondents in large teams in this country.



3.2.8 Interactive Visualisation 3

Angle: How countries perform on large and small data teams?

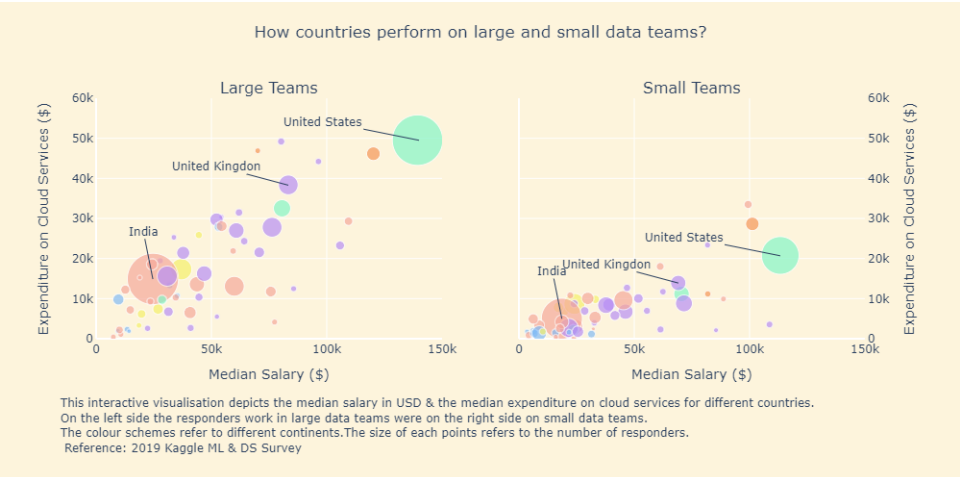
Framing: We generated the following five distinctive categories regarding respondents' spending on cloud services.

	Spending on cloud services	Count
Large	$0 (USD)	1621
	$1-$99	555
	$100-$999	1022
	$1000-$9,999	1008
	$10,000-$99,999	757
	> $100,000	784
Small	$0 (USD)	1434
	$1-$99	711
	$100-$999	1056
	$1000-$9,999	985
	$10,000-$99,999	465
	> $100,000	204

Focus: We intended to find the relationship between average spending on cloud services and the median salary among respondents in different countries.

Annotation: We selected three countries as show their relative position across the large and small teams. And we use different colour to represent continent; we use size of bubble to represent the number of responders.

Interactivity: Viewers can zoom in to have a close view of any specific bubble (represents a country), and they can click the bubble to see the corresponding expenditure on cloud services and the median salary of the country.



3.2.9 Interactive Visualisation 4

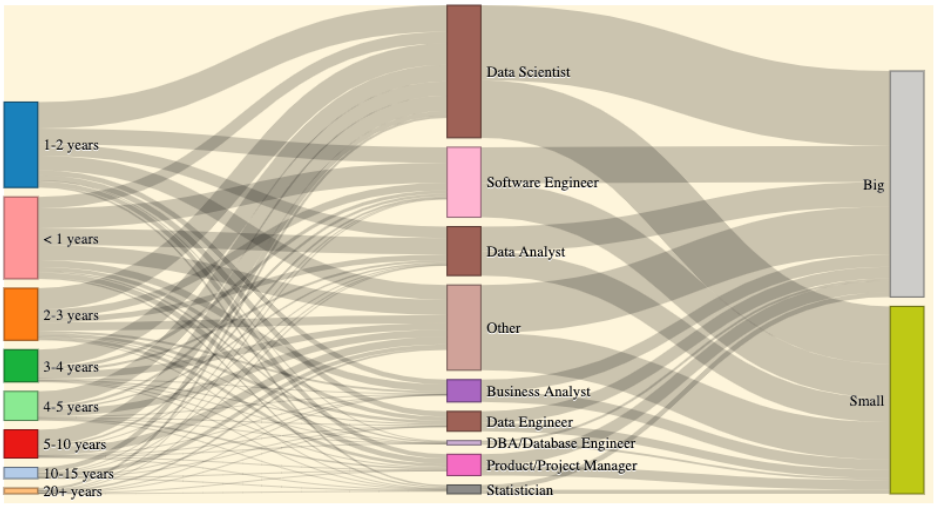
Angle: What are the compositions of the two size of teams and each data-related role's working experience?

Framing: We included all respondents who provided information regarding their role.

Focus: Our goal was to give a sense of the distribution of the responders across different roles, team sizes and year of experience.

Annotation: Though there is no text annotation in this chart, viewers can have a clear idea of (1) what are the compositions of large and small teams respectively; (2) what are the average working experience of each role by perceiving the width of the arcs.

Interactivity: Viewers can click on one specific arc and see the flow of the number of respondents.



3.2.10 Interactive Visualisation 5

This visualisation has been used in the introductory part of the interactive presentation.

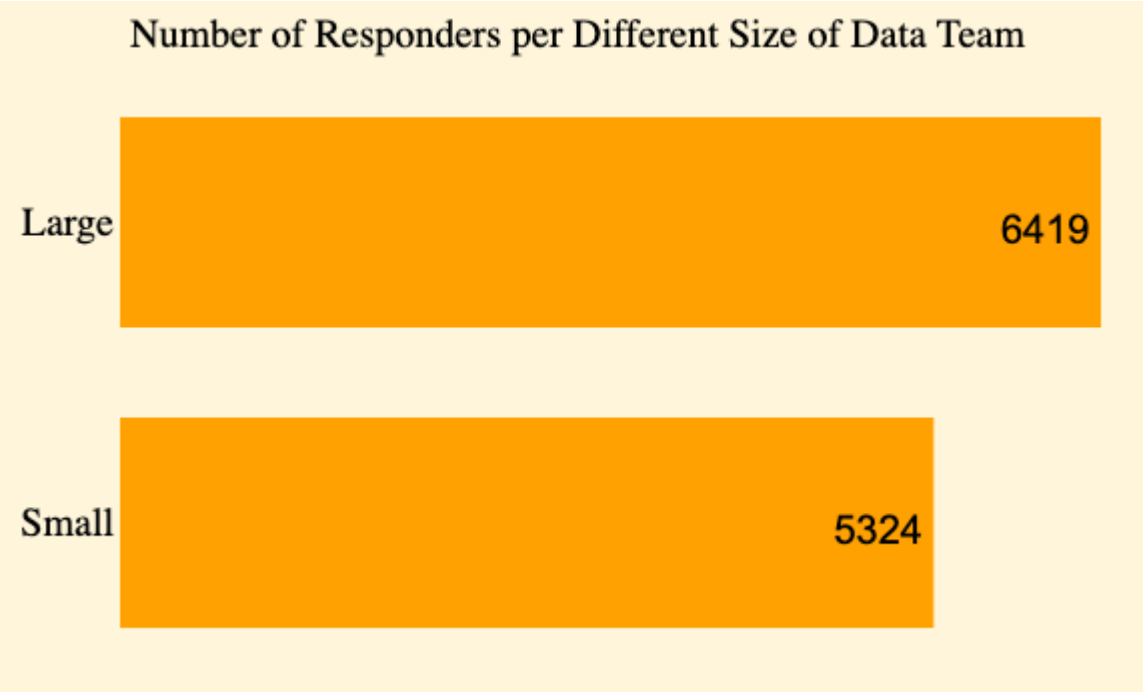
Angle: What is the number of responders?

Framing: We kept only the responders that we have dealt with in our project

Focus: Provide the actual number of responders so viewers can realize the actual size of the survey (there might be other surveys but not with so many responders).

Annotation: Show with a fade in effect the number for each group.

Interactivity: Viewers cannot interact with the plot, although the plot gradually fades in for 4 seconds once the user enters the presentation.



References

1. Kirk, A., 2016. Data Visualisation. 1st ed. Sage.
2. Kaggle.com. 2020. *2019 Kaggle ML & DS Survey*. [online] Available at: <https://www.kaggle.com/c/kaggle-survey-2019> [Accessed 11 June 2020].
3. Dale, K., 2016. Data Visualization With Python And Javascript. O'Reilly Media.
4. Kelleher, C. and Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. Environmental Modelling & Software, [online] 26(6), pp.822-827. Available at: <https://doi.org/10.1016/j.envsoft.2010.12.006>
5. MRW, 2018. Global Data Visualization Market - Segmented by Organizational Department, Delivery Mode, Industry Vertical (BFSI, IT & Telecommunication, Retail/e-commerce, Education, Manufacturing, Government) and Region - Growth, Trends and Forecasts (2018 - 2023). [online] MRW. Available at: <https://www.marketreportsworld.com/global-data-visualization-market-12343651>.
6. Color Palette by Adobe. 2020. [online] Available at: <https://color.adobe.com/create/color-wheel> [Accessed 11 June 2020].

Appendix

Appendix.1 Timeline of the project

Timeline	
25th of May	Selection of Topic
28th of May	Completion of preliminary analysis
8th of June	Completion of solution development
9th of June	Reconciliation of the solution and the analysis
10th of June	Proof-reading – Submission

Appendix.2 Dataset selection (Phase 1)

Dataset Selection Process (Phase 1)

Name	TMDB 5000 Movie Dataset	2019 Kaggle ML & DS Survey	Global Terrorism Database	Lending Club Loan Data
Strengths	1. Easy to explain our motivation 2. Contain many variables 3. contains full credits for both the cast and the crew 4. Can have business impact 5. Can be connected to current situation 6. 38mb with information for which actor has played in each movie (however it will need preprocessing) 7. Records from very old movies; can demonstrate how movie industry has changes (e.g. how many movies were released (cumulative) across years for each category 8. It contains imDB ratings; can show audience preference 9. It contains budget / profits information	1. 3 years data 2. The topic is pretty novel and interesting 3. Have geographical data (Responders across different countries) 4. Related to our degree 5. The results have motivations to be visualized 6. Contain many categorical attributes 7. Many different users (male/female - analysts/scientists - self-educated/uni-educates - inexperienced/experienced - R users/Python - lowpaid-highpaid (needs adjustment as they are from diff countries) users and much more) 8. The first point can actually lead to hierarchies; e.g. analysts-->uni-educated-->inexperience-->Python users 9. Except the previous year's surveys (which they require to be merged) - we have also age attribute; this can show a time range. Eg young people tend to be self-educated where old people are uni-educated? I can image this also as a trend graph	1. Enough Columns 2. Have geographical data 3. Interesting Topic 4. A wide range of data period 5. Have 100+ attributes 6. cleaned file (only 1 csv file, easy to do exploration) 7. We have Perpetrators nationality and target nationality - can show relation between them. 8. Several categorical data; e.g. attack types / target types	1. Have enough variables 2. interesting topic and it is a relatively new filed 3. Have many attributes 4. Contains data from 2007 - 2015, can analyse the trend over years 5. Have good connection to current situation 6. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections among others. 7. Many transactions --> Aggregations (min/max/mean interest for example) 8. Different applications for loans (cars, houses etc) / for 3 or 5 years 9. Interest Rate by different creditability Rate
Weaknesses	1. Trite topic 2. Don't have many categorical variables 3. maybe outdated (its created 3 years ago) 4. Only 5mb of actual data; 20 columns only - Actually some of them are long descriptions (strings) --> NLP --> too much for this project 5. No geographical data	1. Not wide time range (2017-2019). Maybe Age can play the role of time	1. Maybe not easy to explain 2. Hard to find business value 6. Maybe needs domain knowledge (e.g. Geo-political Global situation). Great current and past enemies. 7. there is no official resource for some of the values	1. No geographical data (only zip code) 2. The nature of this type of business may provide a weak justification for the motivations of creating visualization (ML classification models may have more business value) 3. Needs groupbys - moderate calculations 4.More technical topic"
Big Titles (common editorial Theme)	What makes a movie profitable? Does a movie make an actor popular or an actor a movie? (needs more clarification) How sharp is the divide between major film studios and the independents?	What makes someone a Data Analyst or Data Scientist Which skills a Data Scientist should get? (needs elaboration - maybe needs external resources) What factors affect people's choice on studying data science?	Internal Terrorism vs External Terrorism? What is the trend of terrorism over in last 30 years?	How to minimize risk for loan investment? P2P - A new way to take short loans (exploratory across all dataset - not specific question)
Criteria 1: Technical Requirements	×	√	x	x
Criteria 2: Story-telling Requirements	√	√	√	√
Criteria 3: Auidence Profiling Requirement	√	√	x	x

Technical Requirements : The dataset should include many attributes (>30)- this will allow us to create different graphs and focus on different aspects. Ideally it should include geolocation data (geographic areas) & repeat instances (rows) across different time periods (ideally to have a datestamp). This

Story-telling Requirements: Intended Audience? Who will be intrested in reading our 2-A4 page? In which magazine you could include it?

Auidence Profiling Requirement: Does the nature of the dataset have a relatively good amount of user base? Does the stakeholders of the dataset have strong desire of perceiving information from visualization?

Appendix.3 Final Static Presentation

OCULART VISUAL

VOL. JUNE 2020

Kaggle DS&ML Survey 2019

You live for crunching data?

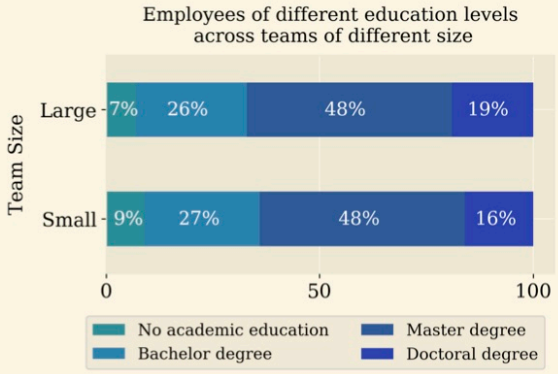
HERE'S WHAT YOU NEED TO KNOW ABOUT THE SIZE OF YOUR TEAM!

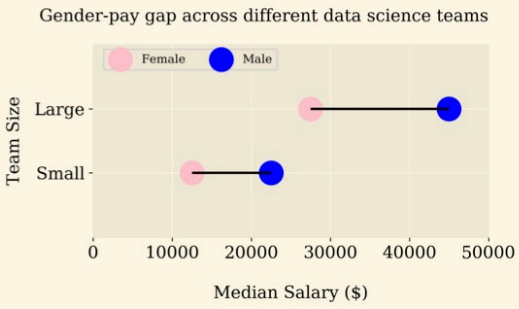
There are two sides regarding the size of the Data Science teams. Small companies are great environment to share ideas, knowledge, be flexible and probably to have better working conditions. On the other hand, big teams may work with bigger projects and there is always the right person for a new task. In this analysis, offered to you by Oculart, you will know more about the differences between small and big data teams. For this purpose, we will review the latest **Kaggle** survey from 11743 professional who work with data. From this survey 5324 responders work in team of 2 to 4 members and 6419 responders work to team of more than 4 members. The data were collected in 2019 from users all over the world.

Kaggle is a Data Science community where data scientist can publish datasets and build machine learning models. It is also a platform for data scientists to work with others to solve data science challenges.

EDUCATION LEVEL

It is true that data teams may come with people from different academic backgrounds. Nowadays, the variety of Massive Open Online Courses (MooC) allows anyone to gain data analytical skills without the need to pursue postgraduate (and probably expensive) studies. From our analysis it seems that still the majority of data professionals have at least a master degree. However, there is no great difference in educational level across small and large teams.



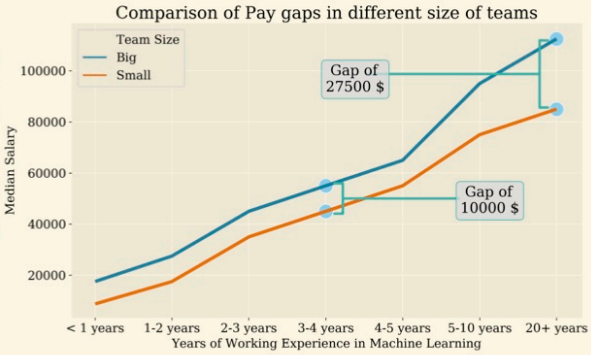


GENDER PAY GAP

Gender pay gap is now a hot topic. It is widely acknowledged that gender pay gap exists in many industries, so how about data related roles? As we can see in the plot, female data specialists tend to have a lower salary compared to male in both team size groups; however, this distance gets wider for respondents in large groups. It is clear that the gender pay gap it is greater in large teams, indicating that gender inequality might be more widespread in large data teams.

WORKING EXPERIENCE IN ML

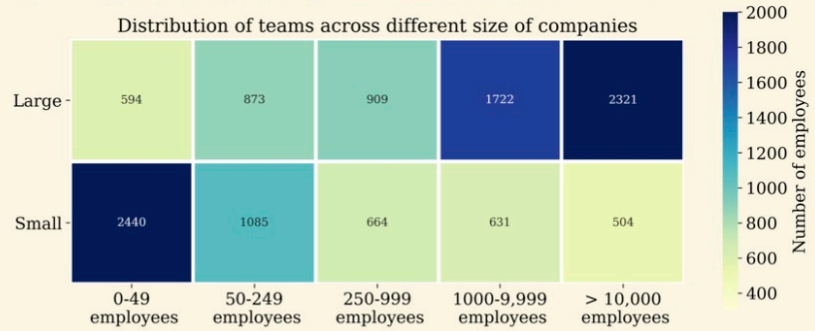
Many people might be also interested in the actual pay gap across as their career grows. In general, the median salary increases as years of working experience in Machine Learning increase in both groups. Moreover, the median salary in large teams outperforms that in small teams (i.e. around 10.000$ pay gap in 4 years of experience and 27.500$ pay gap in more than 20 years of experience).



COMPANY SIZE

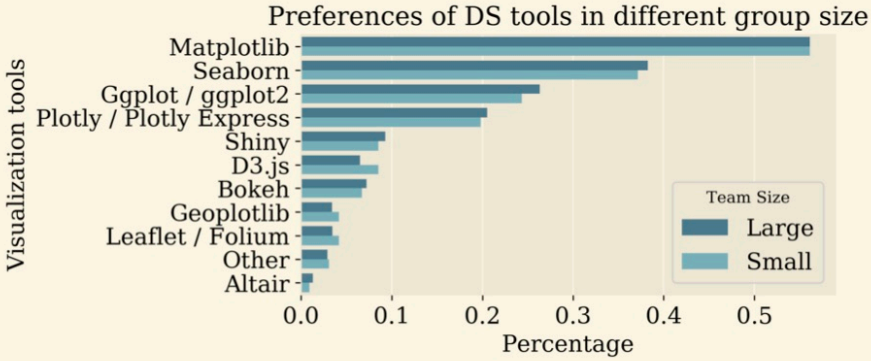
Does the size of a company relate with the size of a data team? Nowadays, more companies trying to exploit the potentials of their data, however not all of the companies have sufficiently large data teams. This heatmap explores the size of data teams for different size of companies. What is interesting is that large data teams may exist also in small teams (probably the most data-oriented ones). On the contrary, we can still find big companies with relatively small data teams (less than 4 people). In the second case, probably the companies are still developing their data teams, or their industry cannot rely to a large extend on data-driven decisions.

Click here to see interactive visualizations!



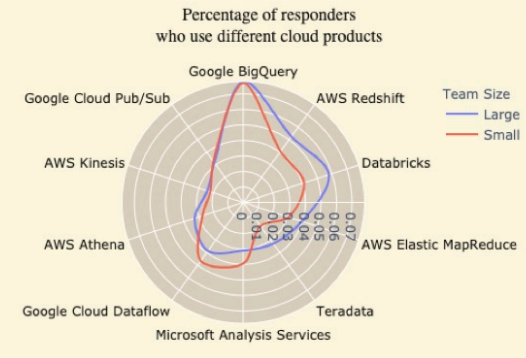
VISUALIZATION TOOL

Visualisation tools play a pivotal role into delivering knowledge to a company. The following plot explores employees' preferences in different visualisation tools. Both groups have almost the same preference of visualization tools. Moreover, it seems that Matplotlib, Seaborn (Python) and ggplot2 (R) are the top three visualization tools used by experts.



CLOUD TOOLS

But what about cloud products? Below we can find the percent of experts who use different cloud frameworks in small and large teams. Overall, it seems that Google BigQuery it's the most popular framework for both small and large teams (7.5%). Large teams show in general a greater preference to many of the cloud products. However, Google Cloud Dataflow and Microsoft Analysis Services are more popular on small teams. Finally, AWS Athena, AWS Kinesis and Google Cloud Pub/Sub have nearly the same popularity for both sizes of team.



Appendix.4 Meeting Notes

OCULART Meeting 1

Meeting App: Microsoft Teams

Date: [20/05/2020]

Time: [12pm-13pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Every group member needs to search for datasets that they find interesting and put some summaries into an Excel worksheet
2. Brainstorming and pick the most voted dataset (create the selection criteria)
3. Explain the expectation/output of the report (Simon)
4. Confirm the framework of our process (e.g. the 4 stages of design)
5. Recommend the textbook (linked provided in Teams)
6. Admin stuff: meeting notes

Action items	Owner(s)	Deadline	Status
Search for dataset Fill in the criteria	[Hailey]	[20/05/2020]	Done
Search for dataset Fill in the criteria	[Symeon]	[20/05/2020]	Done
Search for dataset Fill in the criteria	[Wenhao]	[20/05/2020]	Done

OCULART Meeting 2

Meeting App: Microsoft Teams

Date: [25/05/2020]

Time: [10:30am-11:30am]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Every group member needs to fill the strengths and weakness of the three datasets
2. Choose one of the datasets for the group projects
3. Find and additional aspects from the lecture slides (Lecture 2 and Lecture3 slides)
4. Assigning roles in group work
5. Discuss formulating process (Curiosity, Circumstance, Purpose, Ideas)
6. Discuss Working data process (Acquisition, Examination, Transforming, Exploration)

Action items	Owner(s)	Deadline	Status
- Think and write [Purpose] part as the process of formulating our brief - Examining the dataset (type, size, shape, range)	[Hailey]	[27/05/2020]	Done
- Think and write [Circumstances] part as the process of formulating our brief - Ideas: Check the lecture slides - Find what we should do to clean the dataset	[Symeon]	[27/05/2020]	Done
- Think and write [Curiosity] part as the process of formulating our brief - Set up the selected dataset and rationalize the reason	[Wenhao]	[27/05/2020]	Done

OCULART Meeting 3

Meeting App: Microsoft Teams

Date: [27/05/2020]

Time: [10:30am-11:30am]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Review what we have done so far in the writeup
2. Agreed on how we should proceed with the datasets

3. Assigning roles in group work

Action items	Owner(s)	Deadline	Status
- Exploring the 2019 dataset kernels	[Hailey]	[29/05/2020]	Done
- Merge datasets into one before our exploration	[Symeon]	[29/05/2020]	Done
- Find surveys from other websites - Exploring the 2019 dataset	[Wenhao]	[29/05/2020]	Done

OCULART Meeting 4

Meeting App: Microsoft Teams

Date: [29/05/2020]

Time: [1:00pm-2:00pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Discuss the data cleaning
2. Try to find the interesting indicator of the dataset (Coding language, Country...)
3. Check the dataset of other years (2018, 2017)
4. Discuss the survey questions and answers of the dataset
5. Discuss the procedure of our visualisation project
6. Assigning roles in group work

Action items	Owner(s)	Deadline	Status
- Try to find the interesting point of the dataset for the comparison based on other's notebooks	[Hailey]	[31/05/2020]	Done

Action items	Owner(s)	Deadline	Status
- Try to find the interesting point of the dataset for the comparison based on other's notebooks	[Symeon]	[31/05/2020]	Done
- Try to find the interesting point of the dataset for the comparison based on other's notebooks	[Wenhao]	[31/05/2020]	Done

OCULART Meeting 5

Meeting App: Microsoft Teams

Date: [31/05/2020]

Time: [10:30am-11:30am]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Find the interesting point for the comparison from the dataset
2. Discuss the country related data as our one of the main topics
3. Discuss how to use the ML relevant data and country-related data together
4. Check the survey questions and answers of the dataset
5. Assigning roles in group work

Action items	Owner(s)	Deadline	Status
- Try to do visualisation of the country data - Send an email to the lecturer	[Hailey]	[01/06/2020]	Done
- Try to do visualization of the country data	[Symeon]	[01/06/2020]	Done
- Make a list of EU countries for the country data	[Wenhao]	[01/06/2020]	Done

Action items	Owner(s)	Deadline	Status

OCULART Meeting 6

Meeting App: Microsoft Teams

Date: [01/06/2020]

Time: [13:00pm-14:00pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Make a decision about the project topic
2. Explore the interesting point of the dataset
3. Check the data of gender and company related attributes
4. Prepare the data for visualization
5. Discuss how use the salary dataset for your project
6. Assigning roles in group work

Action items	Owner(s)	Deadline	Status
- Do visualization for the salary dataset	[Hailey]	[02/06/2020]	Done
- Try to do visualization of the country data	[Symeon]	[02/06/2020]	Done
- Do visualization for the company size and gender	[Wenhao]	[02/06/2020]	Done

OCULART Meeting 7

Meeting App: Microsoft Teams

Date: [02/06/2020]

Time: [14:00pm-15:00pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Check the graph what we made (Salary, company size, gender)
2. Discuss how to make better visualization
3. Discuss the better way for coding and visualization
4. Discuss the data pre-processing write up part
5. Assigning roles in group work

Action items	Owner(s)	Deadline	Status
- Write up the data pre-processing	[Hailey]	[04/06/2020]	Done
- Explore different tpyes of visualization - Edit gender pay gap visualization	[Symeon]	[04/06/2020]	Done
- Write up the data pre-processing	[Wenhao]	[04/06/2020]	Done

OCULART Meeting 8

Meeting App: Microsoft Teams

Date: [03/06/2020]

Time: [2pm-4pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Agree on 3 levels of work: notebooks, presentation, writeup
2. Agree on the sequence of the plots:

3. For visualizations that needed to be put first in the presentation, they should be those which capture feelings (those who don't reveal actual statistical numbers)
4. Then our presentation use visualizaitons which focus on numbers

5. The sequence of plots

1st: education level (higher level)

2nd: gender gap

3rd: plot 3B

4th: the heatmap

Action items	Owner(s)	Deadline	Status
Write markdown in the notebook (make a new version of EDA)	[Hailey]	[04/06/2020]	Done
1.edit exisiting plots 2.search for new plots 3.create a doc to do demo of static visualization	[Symeon]	[20/06/2020]	Done
adjust the writeup in the presentation	[Wenhao]	[04/06/2020]	Done

OCULART Meeting 9

Meeting App: Microsoft Teams

Date: [04/06/2020]

Time: [2pm-4pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Discuss how to make Sankey Diagram on the Python

2. Choose visualization method between interactive and statistic
3. Organize the data and visualization for the write up
4. Discuss the labels of the visualization for consistency

Action items	Owner(s)	Deadline	Status
1. Write up about the visualization 2. Finish the writeup of the statistic visualization	[Hailey]	[04/06/2020]	Done
1. Create Sankey Diagram 2. Merge two notebooks (EDA/Annotation) into one notebook	[Symeon]	[05/06/2020]	Done
1. adjust the writeup in the presentation 2. Write markdown in the code	[Wenhao]	[04/06/2020]	Done

OCULART Meeting 10

Meeting App: Microsoft Teams

Date: [05/06/2020]

Time: [1:30pm-3:30pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Discuss the colour of the static presentation background
2. Discuss the overall formatting
3. Organize the data and visualization for the write up
4. Discuss the labels of the visualization for consistency

Action items	Owner(s)	Deadline	Status
Change the formatting of the static presentation	[Hailey]	[07/06/2020]	Done

Add some write up on our static presentation Edit write up Change the colour or labels of the graph for the perfect looking Creating image bar plot	[Symeon]	[07/06/2020]	Done
Change the formatting of the project Edit the size of the visualization for the optimal looking	[Wenhao]	[07/06/2020]	Done

OCULART Meeting 11

Meeting App: Microsoft Teams

Date: [07/06/2020]

Time: [1:30pm-3:30pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Discuss the overall formatting (Colour, Font ..)
2. Organize the data and visualization for the write up
3. Discuss HTML visualization
4. Discuss plotly studio visualization
5. Decided what should we write for the last part of the writing

Action items	Owner(s)	Deadline	Status
Finish writing up	[Hailey]	[08/06/2020]	Done
Make interactive visualization	[Symeon]	[08/06/2020]	Done

Edit the size of the visualization for the optimal looking Finsih Writing up	[Wenhao]	[08/06/2020]	Done

OCULART Meeting 12

Meeting App: Microsoft Teams

Date: [08/06/2020]

Time: [1:30pm-3:30pm]

Attendees: Kim Min Hae (Hailey), Symeon Kokovidis, Wenhao Jiao

Agenda items

1. Discuss the last lecture
2. Interactive visualization adjustment
3. Discuss the write up the description of the statistic graphs
4. Discuss the write up for the interactive visualization

Action items	Owner(s)	Deadline	Status
Write the interactive visualization description	[Hailey]	[08/06/2020]	Done
Make interactive visualization	[Symeon]	[08/06/2020]	Done
Edit Jupyter notebook (write comments in markdown) Write the interactive visualization description (difference between statistic and interactive	[Wenhao]	[08/06/2020]	Done