Data Visualization Assignment 4

Ofir BERGER, Aakash CHAWLA, Hannah HO-LE, Abhimanyu SONI

March 2022

Project Title:

2021 Survey of Kaggle Users.

Link to dashboard:

https://public.tableau.com/shared/42PY4KBS2?:display_count=n&:origin=viz_share_link

Link to demo:

https://youtu.be/tYdSwR0HCaI

Link to github:

https://github.com/SONI-Abhimanyu/Data-Visualization-CentraleSupelec

Introduction

Back in 2012, Harvard Business Review declared: **"Data Scientist: The Sexiest Job of the 21st Century."** Authors Thomas Davenport and D.J. Patil devoted much of their article to defining what a data scientist does. Back then, large enterprises were just waking up to the importance that data science could have in unlocking the power of data. As a relatively new discipline, they noted that the demand for data scientists far exceeded its supply, and "the shortage of data scientists is becoming a serious constraint in some sectors." For data scientists, it's exciting to see salaries, job opportunities, and job satisfaction on the rise. They've certainly earned this as data science has become a key differentiator for many Fortune 500 companies and earned data scientists a seat in the boardroom as critical business decisions are being made.

The dashboard that we develop as a part of this assignment gives an overview of the responses from a survey conducted by Kaggle on its users. The dashboard is targeted for an audience interested in learning more about the demographics of those either in data oriented fields or data enthusiasts, given that Kaggle is essentially a platform for the community of data scientists and machine learning engineers globally. It could be used by students/professionals/researchers etc. who are generally interested in Data Science and similar fields.

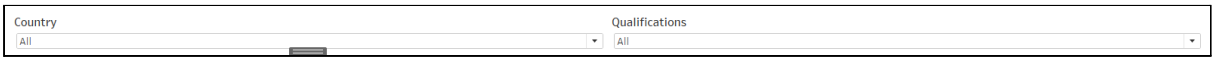
The motivation behind this dashboard is to allow users to gain a quick insight into the 30,000+ responses with just a few clicks and graphs. As future data scientists who are about to graduate, we are very interested in this dashboard as it holds much interest for us as it contains information about the profiles of the data science community and as such can help us frame expectations around pursuing a career in the field, such as how much we can expect to earn, the coding languages or tools we should make sure we are proficient in.

We built the tool using Tableau which is an interactive data visualization software focused on business intelligence. Tableau allows for building interactive dashboards and stories with flexible elements and presents an intuitive interface. We start with building individual graphs which act as foundational building blocks for the dashboard subsequently. Capabilities to filter data on these dashboards were added to enable slicing and dicing in the user customized way.

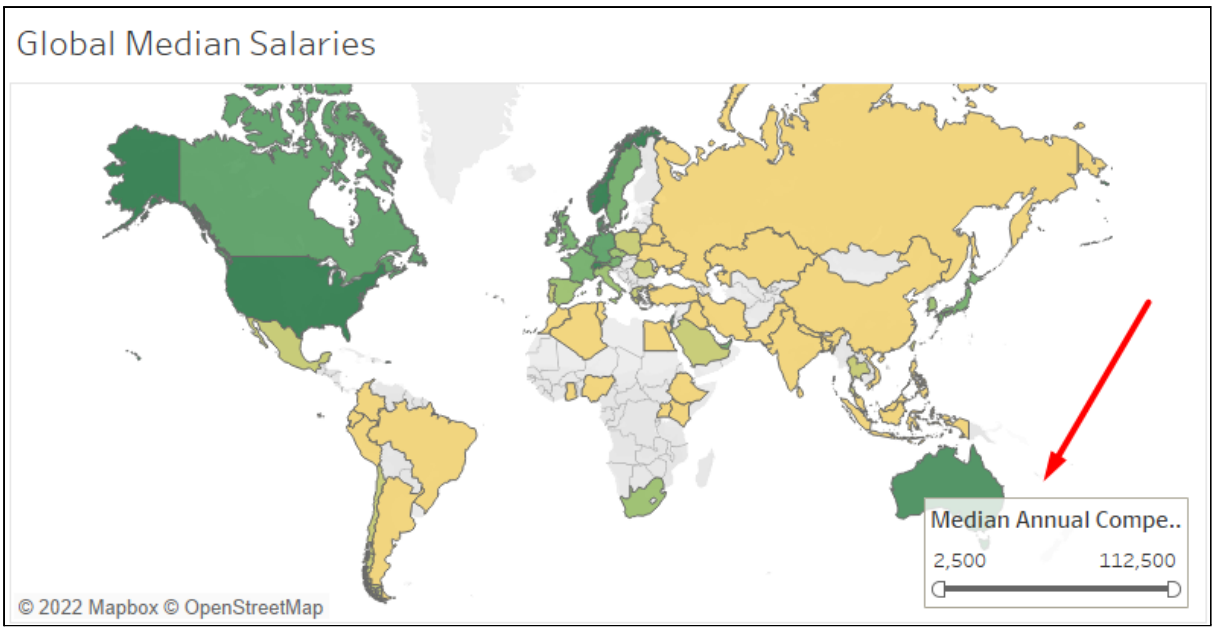
Feature Description

First Tab - Overview

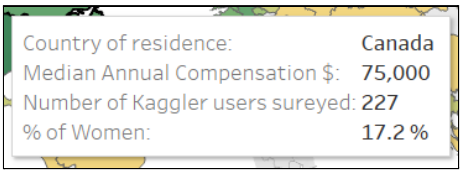
This view gives the user a summary on various important aspects like Employer Industry, Gender, Education Level and Job Title. Users are also provided an added functionality of being able to globally filter all the insights on the page by “Country” and “Qualifications,” there is a default option for “All”, if the user does not wish to apply any filters.



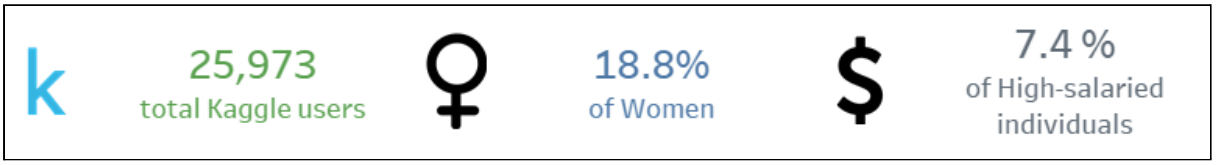
The world map highlights the median annual compensation of the Kaggle Users by their country of origin. A map enables us to present this information in a visually intuitive way by codifying it with colors. It captures attention and illustrates well the difference in median incomes globally, as well as the diversity of respondents in the survey.
The map also acts as a filter and the rest of the visualizations are updated when a particular country is selected either from the map or from the drop down menu. Additionally, there is also a slider feature on the map which gives the user an option to filter the countries based on the median annual salary.



Also, the tooltip on hovering over the countries gives the user access to some useful respective indicators.



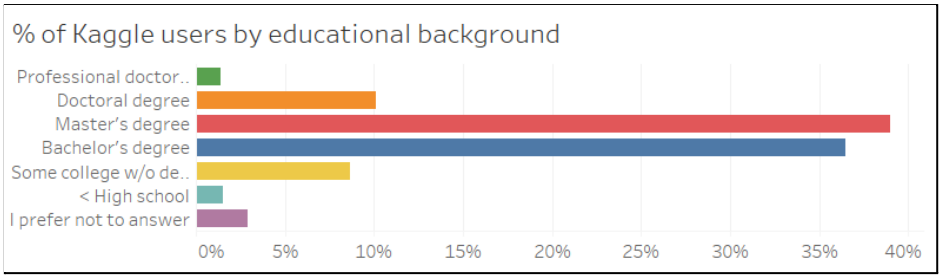
Three KPI's (Key Performance Indicators) on the top right highlight the total number of Kaggle users surveyed, the gender distribution and proportion of high-salaried individuals (earning more than 100,000 USD annually). The KPI's are chosen to be codified as large text to call attention upon them.

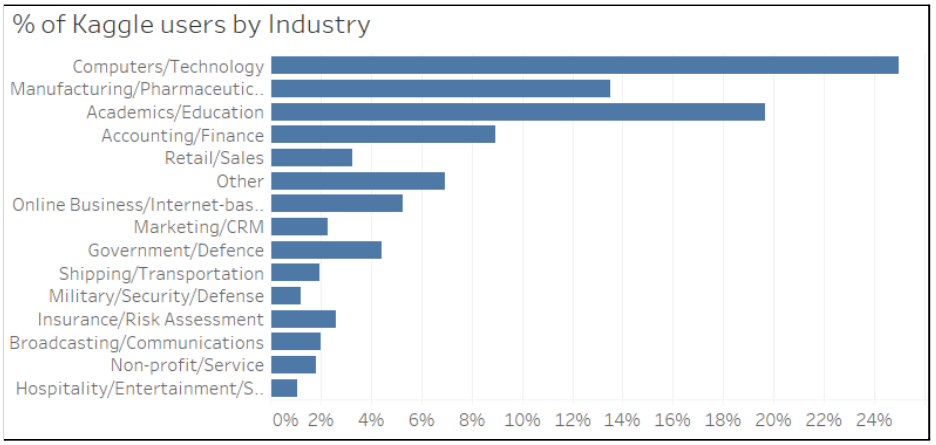


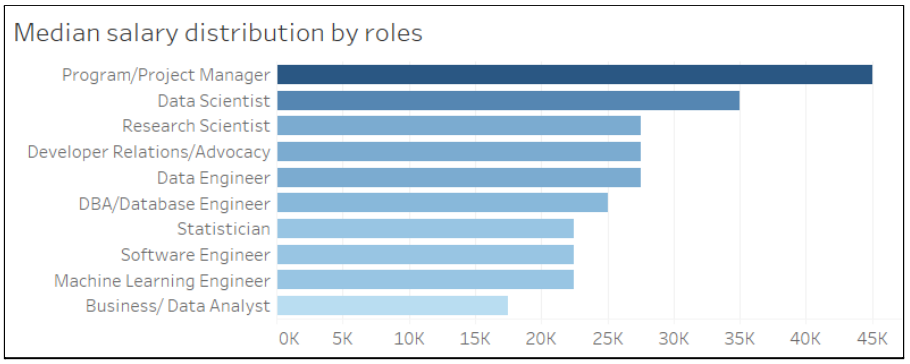
Following three bar graphs presented in this view give an overview of respective distributions for the data that was subset using the filters:

- % of Kaggle users by educational background
- % of Kaggle users by Industry
- Median salary distribution by roles

The bar graphs were chosen since they are very interpretable and convey key information quickly.







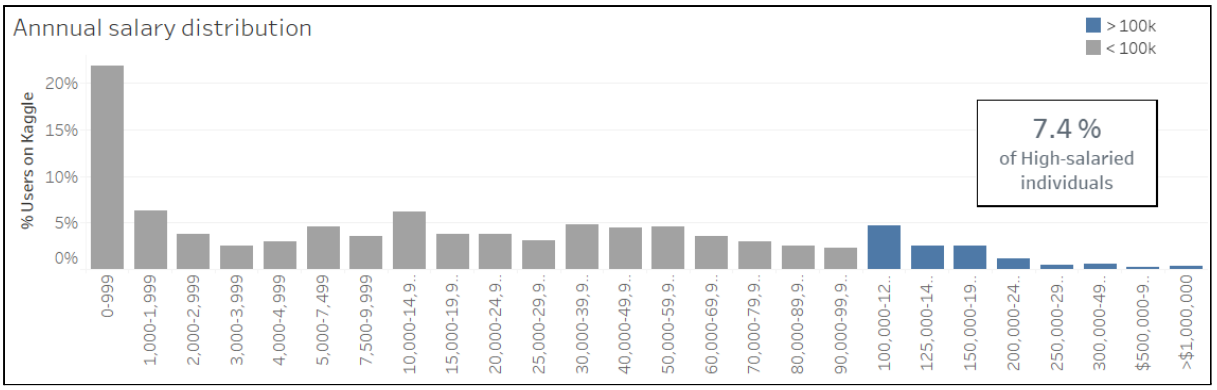
Note that the first of these three graphs, besides presenting information, also act as a filtering for the rest of the view. This visualization is, in fact, bidirectionally linked to the map on the left.

Second Tab - Professional Roles

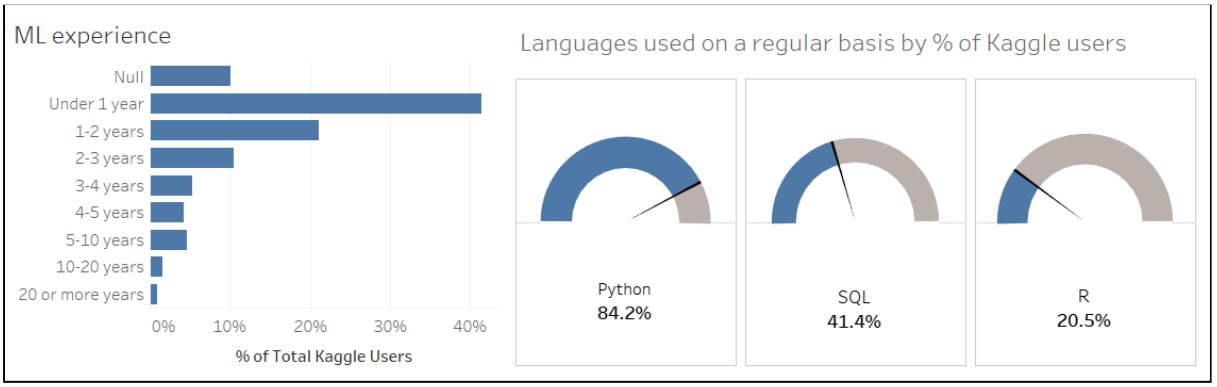
The main features on this tab are the filters for "Professional Role" and "Country", where the user is able to select and get insightful information regarding each professional role and the selected country.
After having explored the industry on surface from the Overview tab, this view gives insights for users with a specific Data Science related professional role in mind.
The provided information is the following:

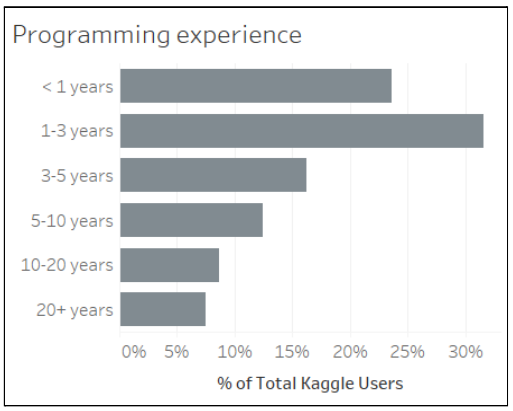
The first visualization highlights the distribution of salary among the individuals for a given Professional Role in a particular geography (as filtered from the top), and emphasizes on the

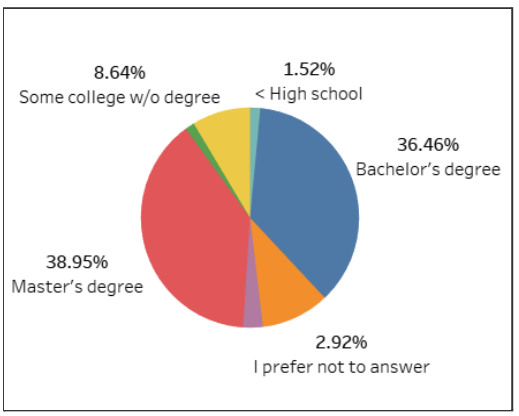
proportion of individuals earning >100K USD annually. A bar chart is an appropriate encoding for a visualization with a discrete distribution.



At the bottom, we bring attention to the experience of individuals in the domain of Machine Learning and programming in general (again these are discrete distributions, hence encoded as bar charts). Proportions of users that use the top 3 programming languages of the domain are highlighted in the form of gauge charts and the distribution of educational backgrounds is encoded as a pie chart. Although these are not the ideal forms of encodings they have been chosen for visual diversity. Respective shares are depicted by the clearly evident percentages.

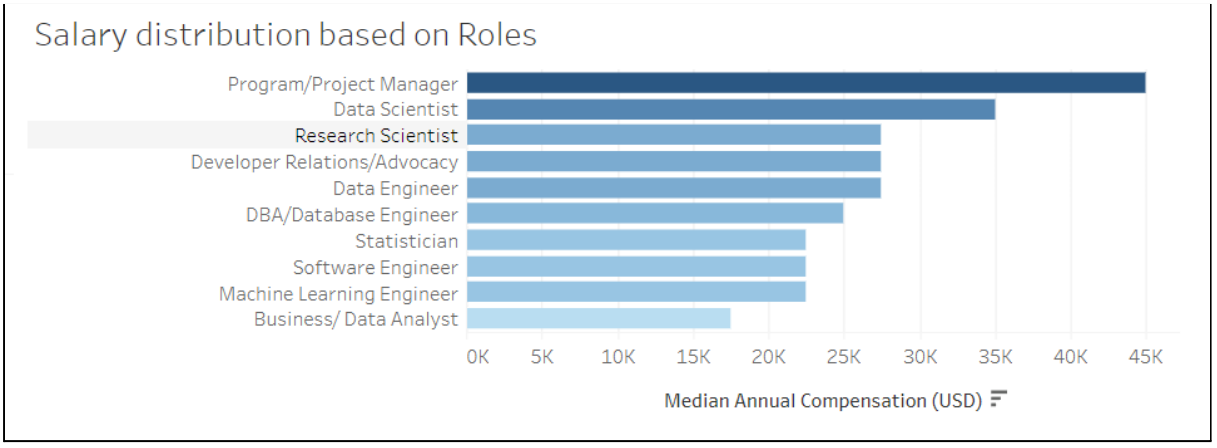




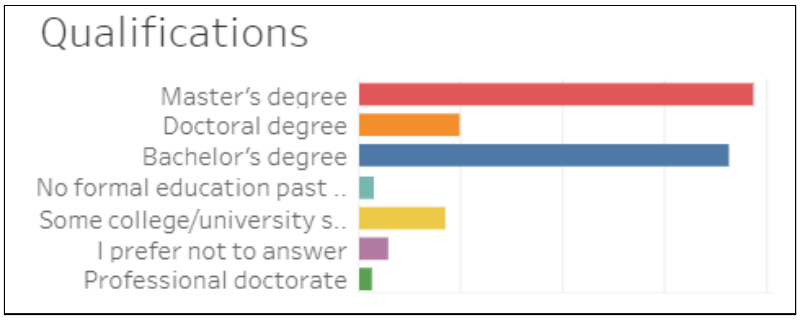


Interesting Findings

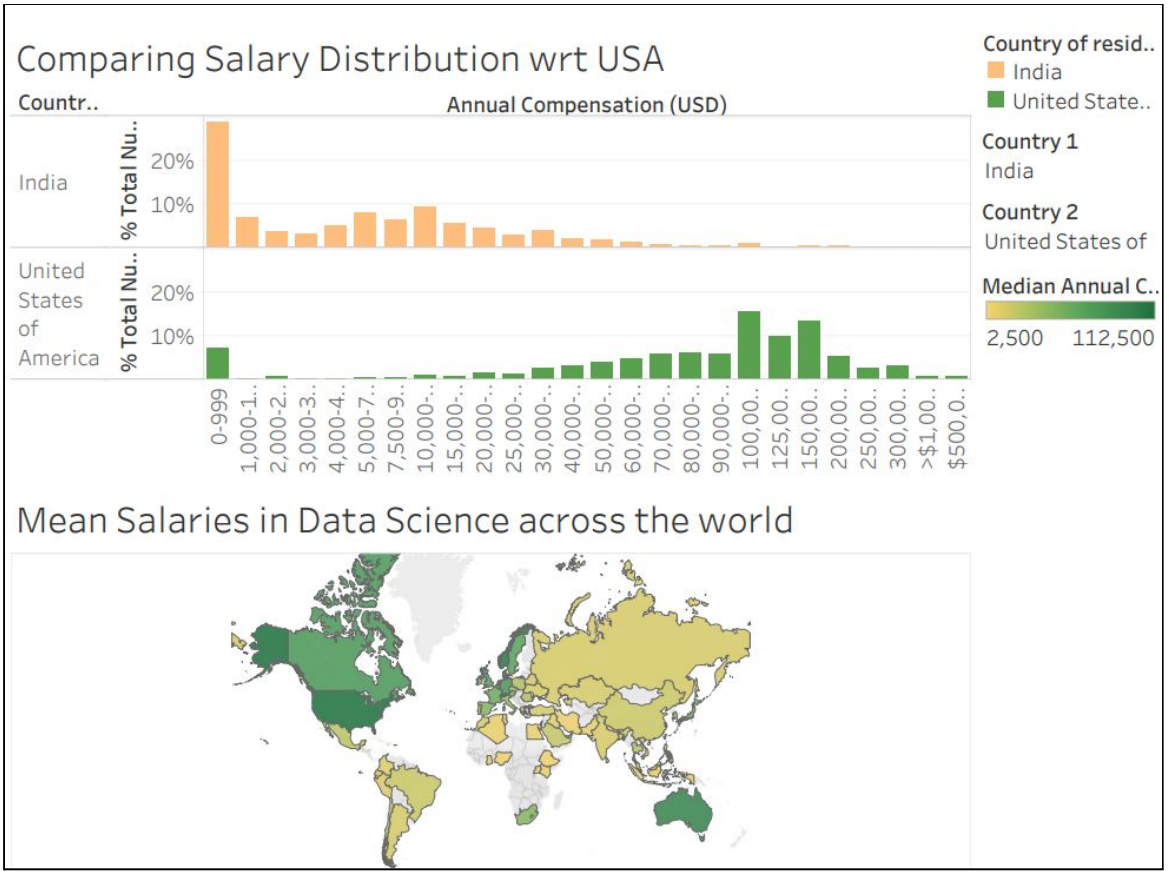
- While Kaggle is a platform targeted predominantly towards data scientists and machine learning engineers, it appears that the most highly paid respondents are program or project managers, earning a median salary of $45k USD.



- Those with a master's degree outnumber those with a bachelor's degree. This is surprising given that Kaggle is used as a learning tool and more people will have a bachelor's degree than a master's. This suggests that perhaps professionals in data related professions are likely to hold a higher master's qualification and thus come to Kaggle to aid them professionally.



- Interesting to compare the salary distribution in data-science/machine learning professionals between India and the USA, as these are the two countries with the largest number of Kaggle users and data-science professionals.
  Median salary in India: $12,500
  Median salary in USA: $112,500
  Median Pay difference between India and the USA: $100,000
  These observations can make us understand why big American companies outsource the data science and engineering jobs to India.



Limitations:

- The insights generated on the dashboard are limited firstly and most importantly by the scope and focus of our project
- The project focuses on understanding and analyzing the different opportunities, salaries and demographics of the labor and workforce in data science industry
- The insights generated on the MAP chart on the dashboard represents nominal amounts of pay or salary, because we wanted to show how absolute salaries vary by geography. A better representation can be adjusting these salaries for inflation and purchasing power parity for each country.
- The dataset has some columns which are very noisy, so there were some limitations on discussing insights from those columns

Contribution of Each Group Member

- All members of the group participated equally in thinking and brainstorming about how our visualization tool would look like.
- Abhimanyu and Hannah - Data Cleaning, Preprocessing, Tableau workbook building and creating the initial version of the visualization tool (dashboard).
- Aakash and Ofir - Additions to the initial tool to refine the storyline and improvements to the visualizations.
- All the group members participated in writing the report.