ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH DỮ LIỆU KHẢO SÁT HỌC MÁY VÀ KHOA HỌC DỮ LIỆU CỦA KAGGLE 2021

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Trần Triệu Vũ	19522539
2	Phạm Đức Thể	19522253
3	Mai Đức Thuận	19522316

TP. HỒ CHÍ MINH – 12/2021

1. GIỚI THIỆU

2021 Kaggle Survey Machine Learning & Data Science là bộ dữ liệu được thu thập, xử lý bởi Kaggle, nội dung các phiếu khảo sát là các khía cạnh xung quanh lĩnh vực Machine Learning & Data Science. Đề tài này sẽ tiến hành phân tích các khía cạnh đó để có những cái nhìn sâu sắc hơn về lĩnh vực này đồng thời đưa ra những định hướng học tập và nghiên cứu cho người mới bước chân vào lĩnh vực này.

Để thực hiện đề tài này nhóm đã sử dụng một số công cụ hỗ trợ xử lý dữ liệu như Pandas, Numpy; công cụ trực quan như Matplotlib, Seaborn; đồng thời kết hợp suy luận từ vốn hiểu biết và các khía cạnh dữ liệu liên quan trong bộ dữ liệu để trả lời các câu hỏi đặt ra từ kết quả trực quan hoặc dẫn dắt phân tích đi đúng hướng.

Thông qua thực hiện đề tài này nhóm đã học được rất nhiều điều đặc biệt là tư duy đặt, giải quyết vấn đề thông qua dữ liệu và có thêm một số hiểu biết nhất định về lĩnh vực mình đang theo học.

2. NỘI DUNG

- Tên bộ dữ liệu: 2021 Kaggle Survey Machine Learning & Data Science.
- Nguồn dữ liệu: Kaggle.
- Cấu trúc gồm: 369 cột là chi tiết của 51 câu hỏi (42 câu hỏi chính và 9 câu hỏi phụ), 25973 dòng dữ liệu, chi tiết:

Câu hỏi số (Số cột)	Cột tương ứng trên dataset	Nội dung
0	Time from Start to Finish (seconds)	Thời gian khảo sát.
1	Q1	Độ tuổi.
2	Q2	Giới tính.
3	Q3	Quốc gia cư trú.
4	Q4	Trình độ học vấn.
5	Q5	Vị trí công việc.
6	Q6	Kinh nghiệm viết code.
7 (13)	Q7_Part1→ Q7_OTHER	Những ngôn ngữ lập trình thông dụng.
8	Q8	Ngôn ngữ mà một data scientist nên học trước tiên.
9 (13)	Q9_Part_1→ Q9_OTHER	Những IDE’s (integrated development environments) thông dụng.

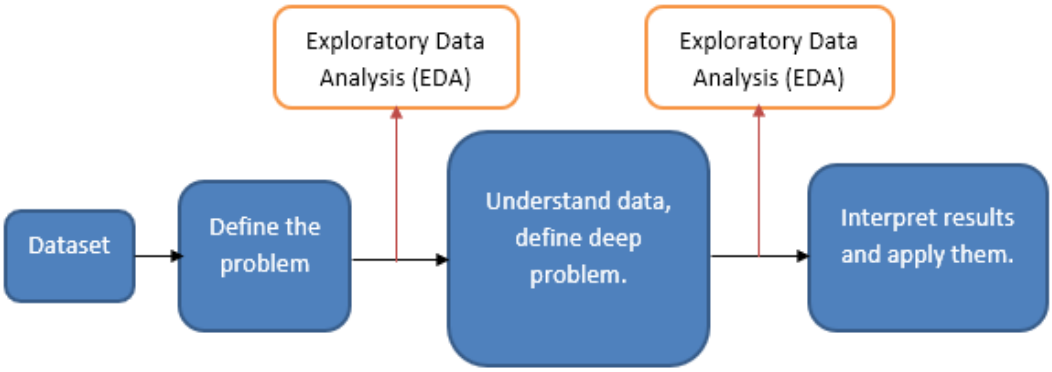
10 (17)	Q10_Part_1→ Q10_OTHER	Những Notebooks thông dụng.
11	Q11	Loại nền tảng tính toán cho các dự án data science.
12 (7)	Q12_Part_1→ Q12_OTHER	Những loại phần cứng chuyên dụng thường được sử dụng.
13	Q13	Số lần sử dụng TPU (tensor processing unit).
14 (12)	Q14_Part_1→ Q14_OTHER	Những thư viện hoặc công cụ trực quan dữ liệu thông dụng.
15	Q15	Kinh nghiệm sử dụng phương pháp machine learning.
16 (18)	Q16_Part1→ Q16_OTHER	Những frameworks thông dụng cho machine learning.
17 (12)	Q17_Part1→ Q17_OTHER	Những thuật toán machine learning thông dụng.
18 (7)	Q18_Part1→ Q18_OTHER	Những phương pháp computer vision thông dụng.
19 (6)	Q19_Part1→ Q19_OTHER	Những phương pháp natural language processing thông dụng.
20	Q20	Chuyên ngành hiện đang làm việc.
21	Q21	Quy mô công ty hiện đang làm việc.
22	Q22	Số người làm về data science trong công ty.
23	Q23	Công ty có sử dụng machine learning hay không?
24 (8)	Q24_Part1→ Q24_OTHER	Những hoạt động là một phần quan trọng trong công việc của người này.
25	Q25	Thưởng hằng năm hiện tại ($).
26	Q26	Chi phí chi cho machine learning, hoặc các dịch vụ điện toán đám mây trong 5 năm vừa qua.
27_A (12)	Q27_A_Part1→ Q27_A_OTHER	Những nền tảng điện toán đám mây thông dụng.
28	Q28	Nền tảng nào cho người này trải nghiệm tốt nhất?

29_A (5)	Q29_A_Part1→ Q29_OTHER	Những sản phẩm điện toán đám mây thông dụng.
30_A (8)	Q30_A_Part1→ Q30_A_OTHER	Những sản phẩm lưu trữ dữ liệu thông dụng.
31_A (10)	Q31_A_Part1→ Q31_A_OTHER	Những sản phẩm quản lý machine learning thông dụng.
32_A (21)	Q32_A_Part1→ Q32_A_OTHER	Những sản phẩm dữ liệu lớn thông dụng.
33	Q33	Sản phẩm dữ liệu lớn tốt nhất.
34_A (17)	Q34_A_Part1→ Q34_A_OTHER	Những công cụ BI (business intelligence) thông dụng.
35	Q35	Công cụ BI người này thường sử dụng nhất.
36_A (8)	Q36_A_Part1→ Q36_A_OTHER	Các loại công cụ machine learning tự động thông dụng.
37_A (8)	Q37_A_Part1→ Q37_A_OTHER	Những công cụ machine learning tự động cụ thể được sử dụng?
38_A (12)	Q38_A_Part1→ Q38_A_OTHER	Những công cụ hỗ trợ người này quản lý các bài viết hoặc sản phẩm machine learning của họ.
39 (10)	Q39_Part1→ Q39_OTHER	Nơi người này thường chia sẻ bài phân tích dữ liệu hoặc ứng dụng machine learning họ làm được.
40 (12)	Q39_Part1→ Q39_OTHER	Nơi người này đã bắt đầu hoặc hoàn thành các khóa học về data science.
41	Q41	Những công cụ phân tích dữ liệu chính.
42 (12)	Q42_Part1→ Q42_OTHER	Các trang mạng xã hội có thảo luận các đề tài về data science.
27_B (12)	Q27_B_Part_1→ Q27_B_OTHER	Nền tảng điện toán đám mây được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
29_B (5)	Q29_B_Part_1→ Q29_B_OTHER	Sản phẩm điện toán đám mây được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
30_B (8)	Q30_B_Part_1→ Q30_B_OTHER	Sản phẩm lưu trữ dữ liệu được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.

31_B (10)	Q31_B_Part_1→ Q31_B_OTHER	Sản phẩm quản lý machine learning được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
32_B (21)	Q32_B_Part_1→ Q32_B_OTHER	Sản phẩm dữ liệu lớn được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
34_B (17)	Q34_B_Part_1→ Q34_B_OTHER	Công cụ BI kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
36_B (8)	Q36_B_Part_1→ Q36_B_OTHER	Loại công cụ machine learning tự động được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
37_B (8)	Q37_B_Part_1→ Q37_B_OTHER	Công cụ machine learning tự động được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.
38_B (12)	Q38_B_Part_1→ Q38_B_OTHER	Công cụ hỗ trợ quản lý các bài viết hoặc sản phẩm machine learning được kỳ vọng sẽ gần gũi hơn với người dùng trong 2 năm tới.

Bảng 1. Thông tin các thuộc tính.

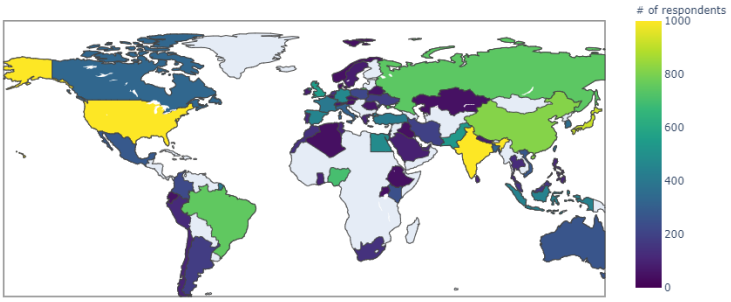
Phương pháp phân tích dữ liệu tập trung vào phân tích thăm dò:



Hình 1. Quy trình phân tích dữ liệu.

Mô tả: Nhóm bắt đầu với việc chọn bộ dữ liệu, kiểm tra các thông tin cơ bản của bộ dữ liệu để đặt ra những vấn đề cơ bản và chung nhất, sau đó tiến hành phân tích thăm dò để giải quyết các vấn đề cơ bản này. Sau khi giải quyết xong nhóm đã hiểu hơn về bộ dữ liệu và phát hiện ra những vấn đề sâu sắc hơn mà bộ dữ liệu mang lại, nhóm tiếp tục tiến hành phân tích thăm dò để giải quyết những vấn đề mới. Cuối cùng nhóm sẽ tổng hợp những kết quả đã phân tích được.

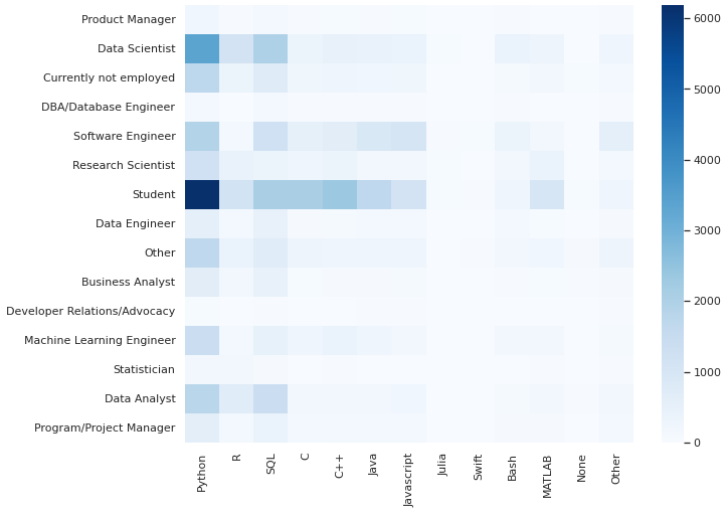
2.1. Những quốc gia nào tham gia khảo sát nhiều nhất?



Hình 2. Số lượng người tham gia khảo sát ở các quốc gia.

Mỹ, Ấn Độ là 2 nước có số lượng người tham gia cao nhất, kế đó là Nhật Bản và Trung Quốc. Nguyên nhân có thể vì các nước này có nền công nghệ được ưu tiên phát triển mạnh nên số người học tập nghiên cứu về lĩnh vực này nhiều hơn hẳn so với các quốc gia còn lại, ngoài ra còn có thể vì Kaggle phổ biến hơn ở các nước trên.

2.2. Ngôn ngữ lập trình nào dành cho bạn?

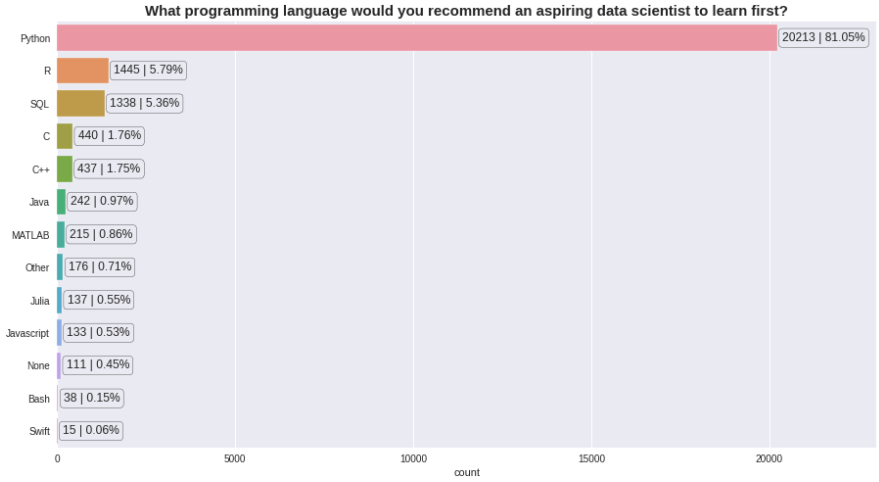


Hình 3. Quan hệ giữa nghề nghiệp và ngôn ngữ lập trình sử dụng.

Ngôn ngữ mà nghề nghiệp bạn hướng đến nên sử dụng là ô có màu đậm nhất hàng:

- Ta không nên xét sinh viên vì kết quả trên chỉ cho thấy sinh viên được học rất nhiều ngôn ngữ để phục vụ cho nhiều định hướng khác nhau của họ.
- Về các vị trí khác, giả sử bạn muốn trở thành một Software Engineer, bạn nên sẵn sàng làm việc với Python, SQL, JavaScript và Java.
- Nếu bạn muốn trở thành một Data Scientist, bạn nên chuẩn bị các kỹ năng cần thiết với Python, SQL, R và Java.
- Nếu bạn muốn trở thành một Data Analyst, bạn nên sẵn sàng làm việc với Python, SQL, R.

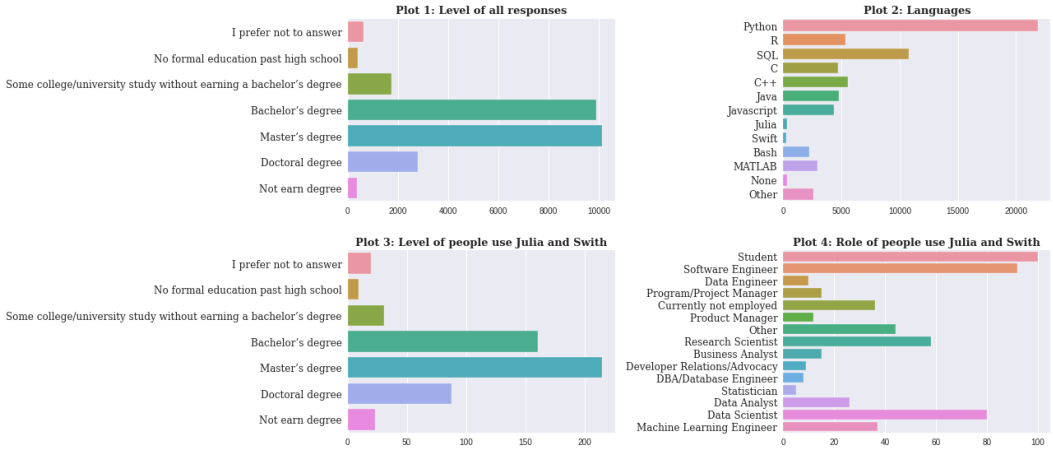
2.3. Những người tham gia khảo sát khuyến khích bạn học ngôn ngữ lập trình nào đầu tiên nếu muốn trở thành một Data Scientist?



Hình 4. Ngôn ngữ lập trình một Data Scientist nên học đầu tiên.

Python là ngôn ngữ được trên 20000 phiếu bình chọn trên tổng số 25973 phiếu (81.05%), đây là một kết quả áp đảo khẳng định sự quan trọng của Python đối với những người có khao khát trở thành một Data Scientist trong tương lai.

2.4. Tại sao Julia và Swift không được sử dụng nhiều?

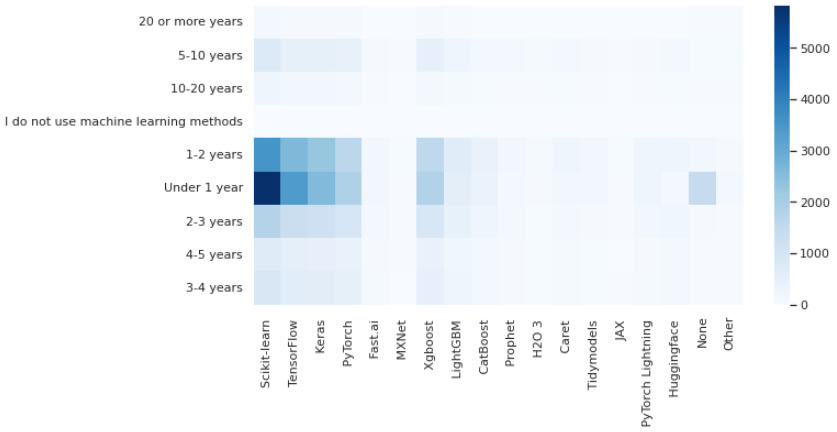


Hình 5. Trình độ học vấn và nghề nghiệp người dùng Julia, Swift.

- Biểu đồ 2 cho thấy Julia, Swift không được sử dụng phổ biến. Có 2 giả thuyết đặt ra là: (i) 2 ngôn ngữ này không phù hợp với chuyên ngành Machine Learning & Data Science; (ii) chúng yêu cầu chuyên môn cao.
- Biểu đồ 4 cho thấy các nghề liên quan đến Machine Learning & Data Science như Data Scientist, Machine Learning, Data Analysis… vẫn sử dụng không ít, chưa thể tin giả thuyết (i).
- Biểu đồ 1 và 3 cho thấy tỷ lệ Master's degree/Bachelor's degree đã tăng từ xấp xỉ 1:1 ở tổng thể lên xấp xỉ 4:3 ở số người dùng Julia, Swift; tương tự với Doctoral degree cũng từ xấp xỉ 1:5 lên xấp xỉ 1:2; vậy có thể khẳng định rằng giả thuyết (ii) đáng tin hơn so với giả thuyết (i), rằng 2 ngôn ngữ này phức tạp để tiếp cận nên mới ít người sử dụng hơn.

2.5. Kinh nghiệm sử dụng các phương pháp Machine Learning của người tham gia khảo sát nói lên điều gì?

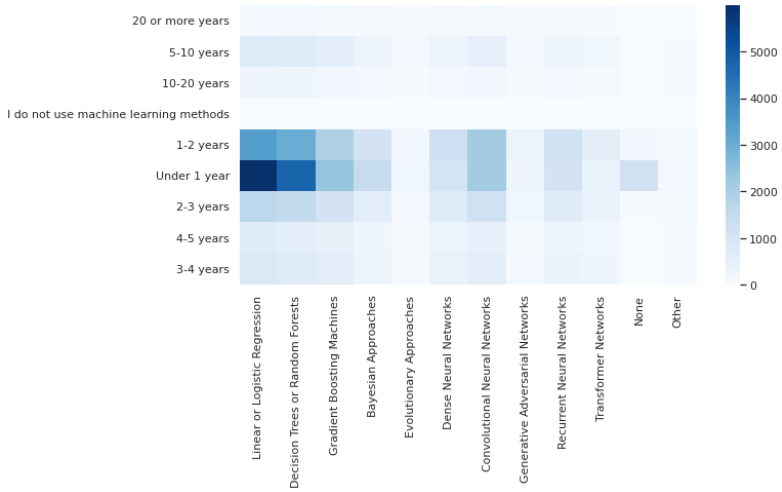
2.5.1. Về những frameworks:



Hình 6. Quan hệ giữa kinh nghiệm và frameworks sử dụng.

- Ta thấy những người đã từng tiếp xúc với các phương pháp học máy dù ít hay nhiều kinh nghiệm đều quan tâm và sử dụng Scikit-learn, Keras, Pytorch, Xgboost, chứng minh sự cần thiết của các frameworks này đối với bất cứ ai làm trong lĩnh vực Machine Learning.
- Có một phần nhỏ sinh viên không biết đến các frameworks này, khả năng là vì với kinh nghiệm dưới 1 năm, họ chỉ mới bắt đầu với các kiến thức hàn lâm và chưa tiến hành thực nghiệm nhiều hoặc có thể họ không học vào chuyên ngành Machine Learning & Data Science.

2.5.2. Về những nhóm thuật toán Machine Learning:



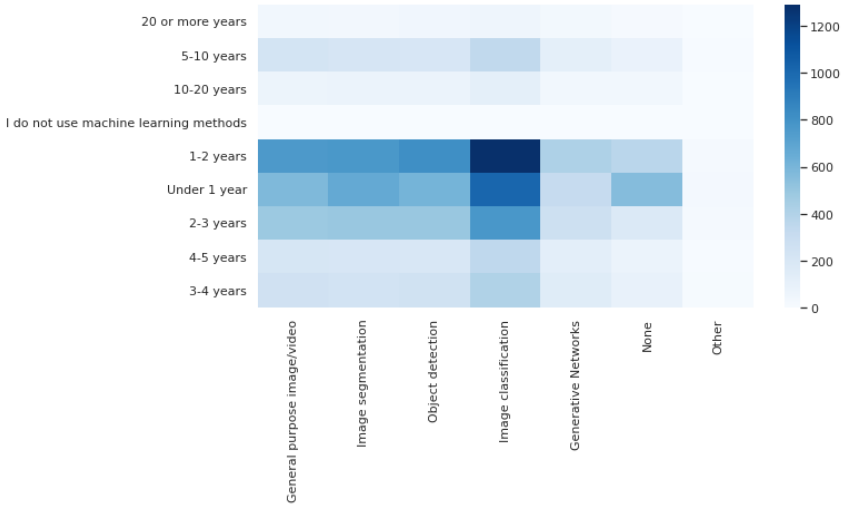
Hình 7. Quan hệ giữa kinh nghiệm và thuật toán học máy sử dụng.

- Linear or Logistic Regression và Decision Trees or Random Forests là 2 nhóm thuật toán cơ bản nhất cho người mới bắt đầu.

- Bên cạnh các thuật toán cơ bản ta thấy được mọi người cũng rất quan tâm đến 2 nhóm thuật toán về Convolutional Neural Networks và Recurrent Neural Networks, tập trung vào những người có kinh nghiệm 3 năm trở xuống. Có thể giải thích cho kết quả như sau: 2 bài toán lớn trong lĩnh vực Machine Learning ở thời điểm hiện tại là Computer Vision và Natural Language Processing, và 2 nhóm thuật toán trên chính là 2 nhóm thuật toán nền tảng tương ứng để giải quyết 2 bài toán này.

2.5.3. Về những bài toán Computer Vision:

- Bộ dữ liệu trình bày các thuật toán theo nhóm, và phân nhóm theo bài toán.
- Ta thấy được bài toán cơ bản nhất trong lĩnh vực Computer Vision là Image Classification, mọi người đều biết qua vì khả năng họ đều bắt đầu với bài toán này. Ngoài ra có thể bài toán hiện còn khó nhất trong lĩnh vực này là các bài toán sử dụng Generative Networks, nó được ít người sử dụng hơn.
- Ở năm đầu tiên ta thấy mọi người quan tâm đến Image Segmentation nhất trong 3 bài toán General purpose image/video, Image Segmentation và Object detection. Nhưng sang năm thứ 2 và 3 số lượng ấy dần chia đều ra, khả năng là ở năm đầu tiên họ thấy đây là một bài toán hay nhưng sau đó họ phát hiện nó còn khó hoặc ứng dụng chưa mạnh họ ít tập trung vào nó hơn ở năm 2 năm 3.

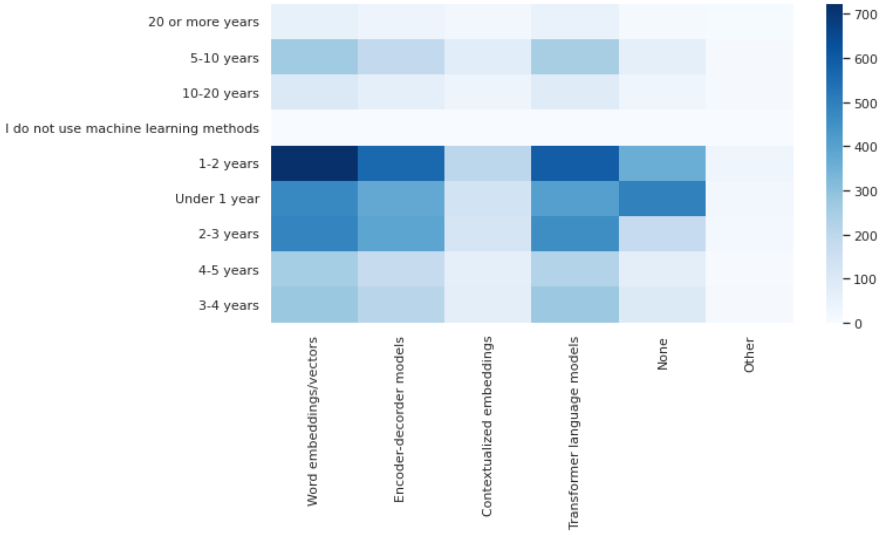


Hình 8. Quan hệ giữa kinh nghiệm và thuật toán Computer Vision.

2.5.4. Về những phương pháp trong Natural Language Processing:

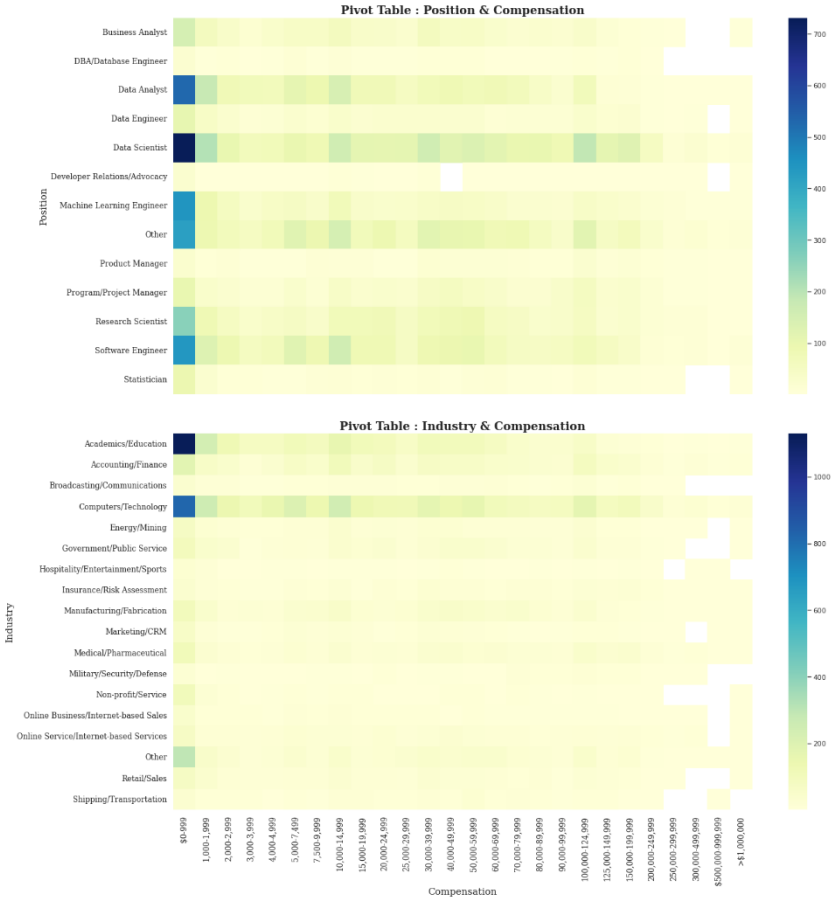
- Các phương pháp được sử dụng xuyên suốt thời gian làm việc của mọi người là kỹ thuật Word embeddings/vector, Encoder - decoder models và Tranformer languages models. 2 kỹ thuật Word embeddings/vector, Encoder - decoder models là cơ sở cho bài toán này và Tranformer languages models là một phiên bản cải tiến trong những năm gần đây.

- Kỹ thuật Contextualized embedding rất ít so với 3 nhóm còn lại vì bản thân nó còn mới và rất khó để sử dụng, kể cả những người có nhiều kinh nghiệm họ vẫn ưu tiên sử dụng 3 kỹ thuật kia hơn.



Hình 9. Quan hệ giữa kinh nghiệm và kỹ thuật trong Natural Language Processing.

2.6. Vị trí nghề nghiệp, lĩnh vực hoạt động ảnh hưởng như thế nào đến mức thưởng hằng năm?

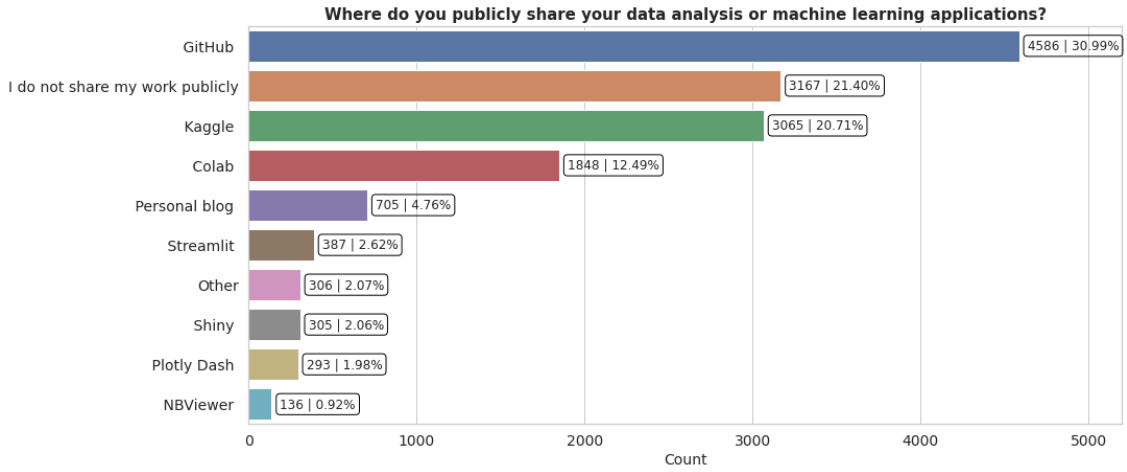


Hình 10. Quan hệ giữa nghề nghiệp, chuyên ngành làm việc và mức thưởng hằng năm.

- Xét về vị trí nghề nghiệp, ta thấy được hầu hết mọi vị trí đều có xu hướng chung là số lượng người giảm dần theo chiều tăng của mức thưởng.
- Các công việc Data Scientist, Data Analyst, Business Analyst có dãy màu ở giữa đậm hơn so với các công việc khác, có nghĩa là số người có mức thưởng tầm trung nhiều hơn. Đây là những công việc đáng để cân nhắc, ngoài ra còn có Software Engineer theo sau 3 vị trí trên.
- Xét về chuyên ngành, các chuyên ngành làm việc chính của mọi người là Academics/Education, Computers/Technology, Accounting/Finance.
- Ngoài ra mức thưởng của 3 nhóm ngành này cũng khá cao đặc biệt là Computers/Technology, dãy màu ở giữa tương ứng với mức thưởng trung bình và cao có màu đậm hơn hẳn.

2.7. Chúng ta có thể tham khảo các bài phân tích dữ liệu, các sản phẩm Machine Learning ở đâu?

- Github, Kaggle, Colab là những nơi phổ biến nhất mà mọi người thường đăng tải thành quả phân tích, nghiên cứu của mình về lĩnh vực Machine Leaning & Data Science, nên đây sẽ là những nguồn tài liệu tham khảo bổ ích cho người muốn tìm hiểu.

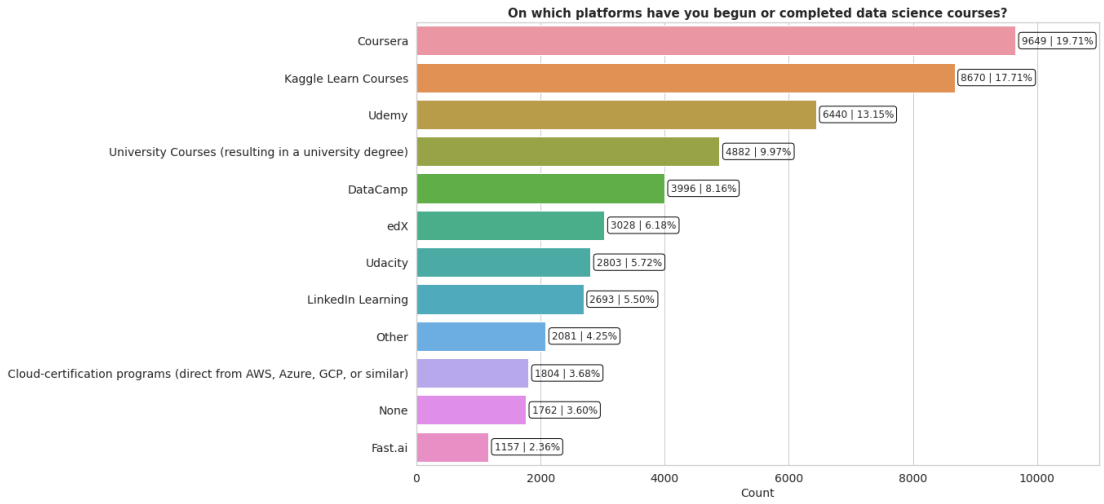


Hình 11. Nơi đăng tải các phân tích, sản phẩm ML&DS.

- Ngoài ra chúng ta cũng nên theo dõi những cá nhân có thành tích nổi bậc trong lĩnh vực này để cập nhật các sản phẩm trên blog cá nhân của họ.

2.8. Nên tham khảo các khóa học ở đâu khi muốn tìm hiểu về Machine Learning & Data Science?

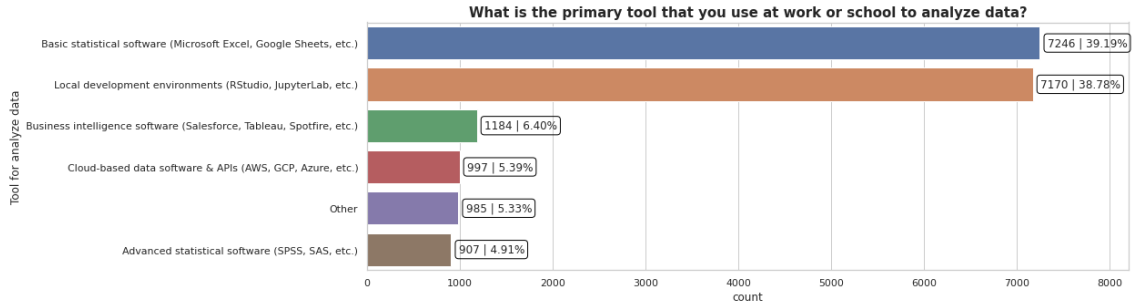
- Phần lớn mọi người đều đã học qua các khóa học trên Coursera và Kaggle Learn Courses, Udemy, đây là những nền tảng nổi bậc, kế đó là các khóa học tại các trường Đại học. Đây là những nơi nên ưu tiên đăng ký học, có thể nhiều người học vì các khóa học ở đây chất lượng, uy tín, cũng có thể vì một số trong đó miễn phí.
- Ngoài ra mọi người cũng tham khảo và học thêm ở các nền tảng khác như DataCamp, edX, Udacity, Linkedln learning,…



Hình 12. Nơi mà người tham gia khảo sát học Machine Learning & Data Science.

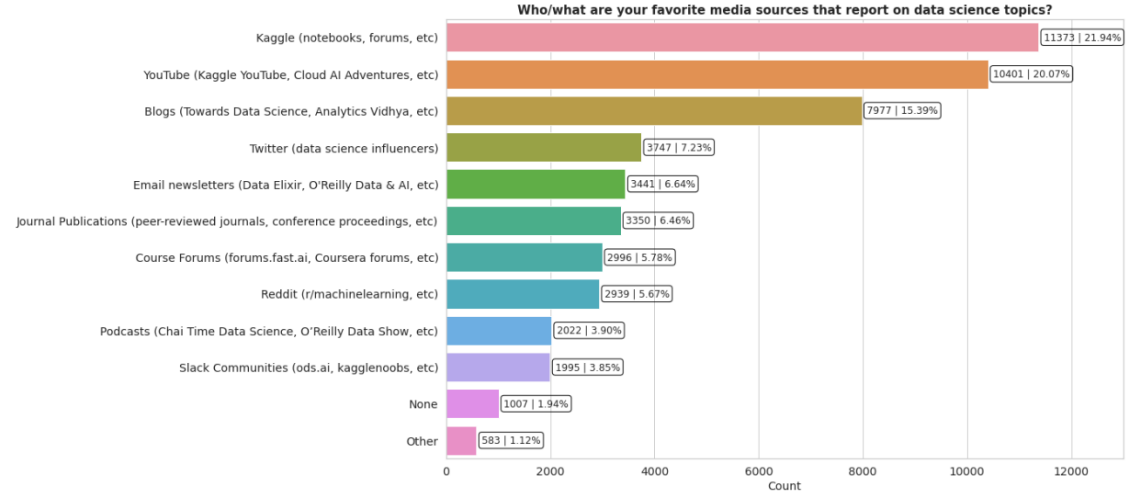
2.9. Bạn nên làm quen dần với các công cụ phân tích dữ liệu nào?

- Mọi người đều ưu tiên dành sự quan tâm cho 2 nhóm công cụ thống kê và phân tích dữ liệu là Basic statistical software như Excel, Google Sheets,… và Local development envirements như Rstudio, JupyterLab,… Đây là những công cụ cơ sở nhất cho việc phân tích dữ liệu.



Hình 13. Công cụ phân tích dữ liệu phổ biến.

2.10. Bạn nên cập nhật các chủ đề về Data Science từ các nguồn nào?



Hình 14. Các trang mạng xã hội có chủ đề về Data Science.

- Các mạng xã hội để cập nhật thông tin về Data Science của mọi người rất đa dạng, có thể vì môi trường làm việc khác nhau, khu vực cư trú khác nhau nên các trang mạng xã hội gần gũi với họ cũng khác nhau.
- Mọi người thường chọn Kaggle; Youtube; các Blogs như Towards Data Science, Analytics Vihya,… và Twitter để cập nhập các chủ đề về Data Science. Đây là những nguồn truyền thông miễn phí, nhiều người cập nhật báo cáo nên được nhiều người tham gia khảo sát ưu tiên quan tâm.

3. KẾT LUẬN

Trong suốt quá trình làm việc, nhóm đã hoàn tất việc EDA trên tất cả các thuộc tính của bộ dữ liệu, tìm hiểu và phân tích sâu hơn nhiều vấn đề được đặt ra sau khi EDA cơ bản bộ dữ liệu. Trên đây là những kết quả được chọn lọc từ những phân tích của nhóm, toàn bộ các phân tích sẽ được trình bày trong source code.

Thông qua quá trình phân tích nhóm đã trả lời được rất nhiều câu hỏi liên quan về lĩnh vực Machine Learning & Data Science như Những quốc gia nào tham gia khảo sát nhiều nhất? Ngôn ngữ lập trình nào dành cho bạn? Những người tham gia khảo sát khuyến khích bạn học ngôn ngữ lập trình nào đầu tiên nếu muốn trở thành một Data Scientist?… Đây là những giá trị quan trọng mà bộ dữ liệu mang lại.

TÀI LIỆU THAM KHẢO

[1] Kaggle. kaggle.com/c/kaggle-survey-2021/overview (28/11/2021).

[2] plotly | Graphing Libraries. plotly.com/python/choropleth-maps/ (19/11/2021).

[3] Seaborn. seaborn.pydata.org/generated/seaborn.heatmap.html (22/11/2021).

[4] Seaborn. seaborn.pydata.org/tutorial/color_palettes.html (22/11/2021).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Trần Triệu Vũ	- Phân tích tổng quan trên dữ liệu. - Tham gia báo cáo hằng tuần và chọn các phân tích hay đưa vào báo cáo. - Tổng hợp các kết quả vào báo cáo. - Tham gia sửa báo cáo, source code, slide.
2	Phạm Đức Thể	- Phân tích tổng quan trên dữ liệu. - Tham gia báo cáo hằng tuần và chọn các phân tích hay đưa vào báo cáo. - Tổng hợp souce code từ 3 thành viên vào 1 file. - Tham gia sửa báo cáo, source code, slide.
3	Mai Đức Thuận	- Phân tích tổng quan trên dữ liệu. - Tham gia báo cáo hằng tuần và chọn các phân tích hay đưa vào báo cáo. - Trình bày slide. - Tham gia sửa báo cáo, source code, slide.