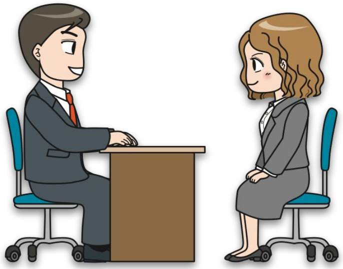
Sabancı Üniversitesi

Job Satisfaction of Kagglers

An Analysis On Education & Job Data of Kaggle members

B.A.R (Group 21)

Aleeze Abbas
Raneem Elshanawany
Nur Başak Özer



Kaggle Notebook: https://colab.research.google.com/drive/1JUDYbPiZAoOkJ8M43XfkhB6zhP6ei7_2?usp=sharing

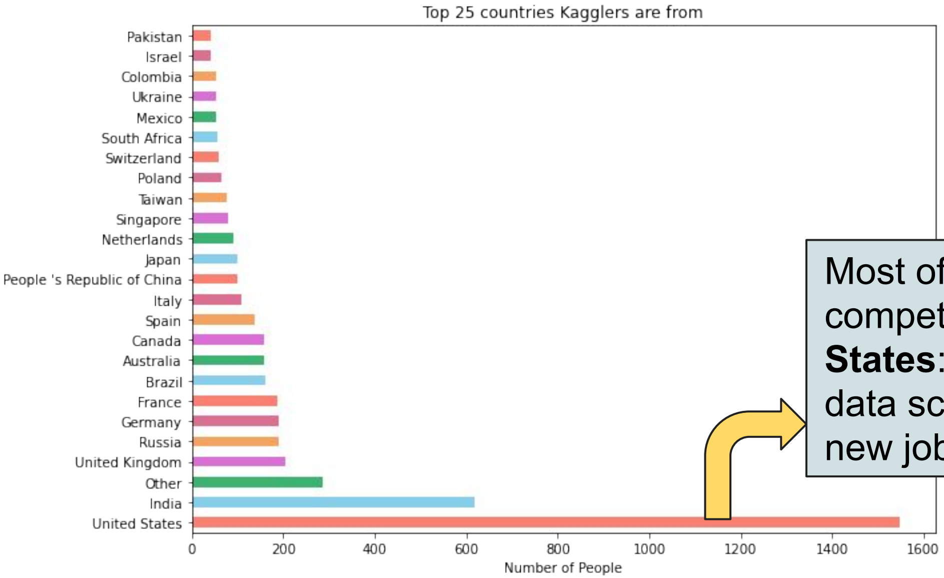
Introduction to the world of kaggle

- Kaggle is an online community of data scientists and machine learning practitioners with **over 1 million registered users** as of June 2017.
- Kaggle has run hundreds of machine learning competitions which have resulted in many successful projects including the state of art in HIV research, chess ratings and traffic forecasting.
- Thus, the data we have collected from Kaggle competition participants is highly beneficial for our up-and-coming learning platform **IMLearning** which hopes to attract many data scientists and machine learning engineers from all around the world.

Kaggle: a diverse community

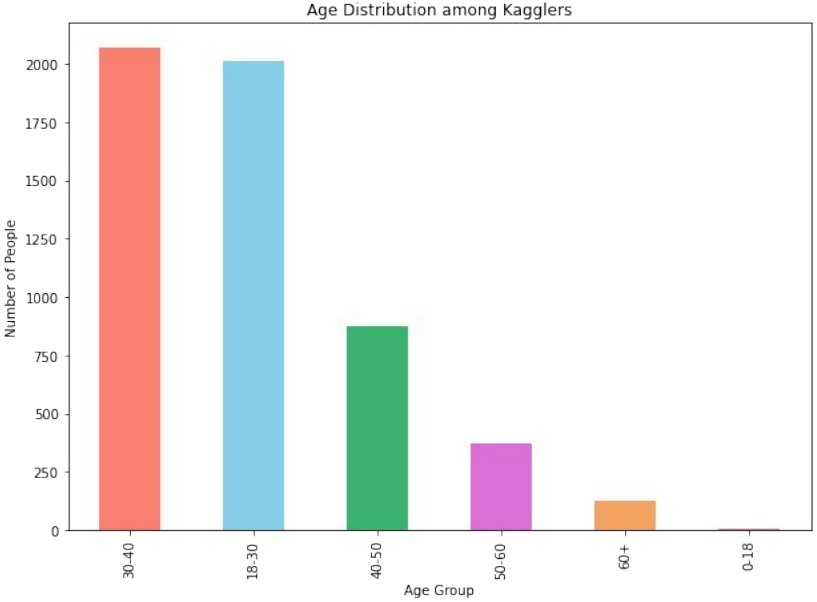
❖ From a business standpoint, one of main goals that **IMLearning** wants to achieve is to attract as many data scientist as possible. For that we have performed a demographic analysis on the Kagglers to have a grasp of our potential customers.

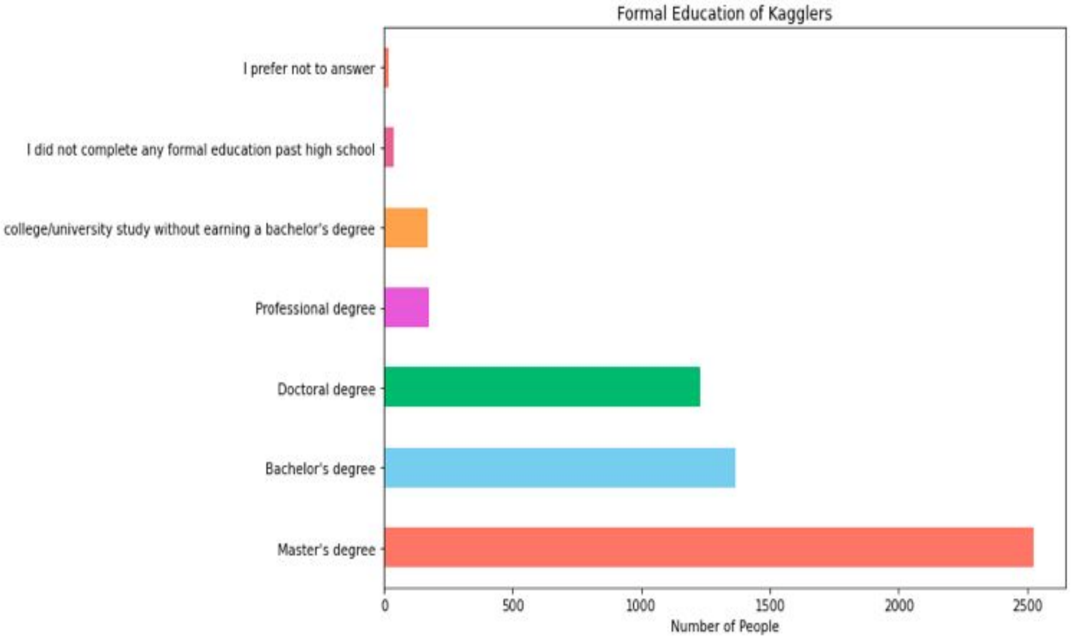
However, it confirms the fact that Kaggle is a highly diverse community which spans 194 countries.



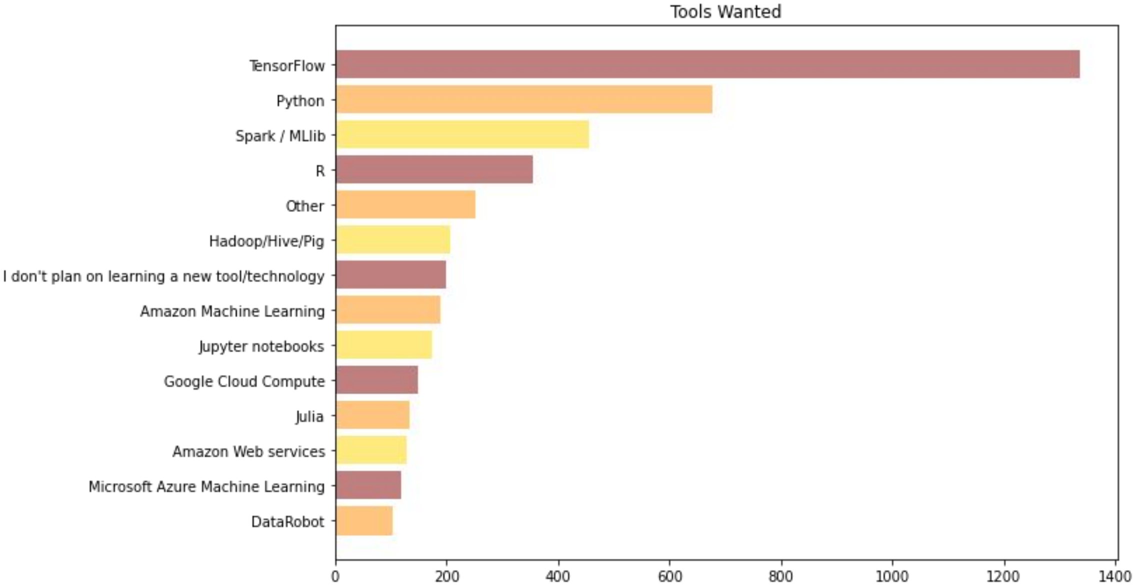
Most of the participants of Kaggle competitions were from the **United States**: an appropriate result since data science has been deemed "the new job in the United States"

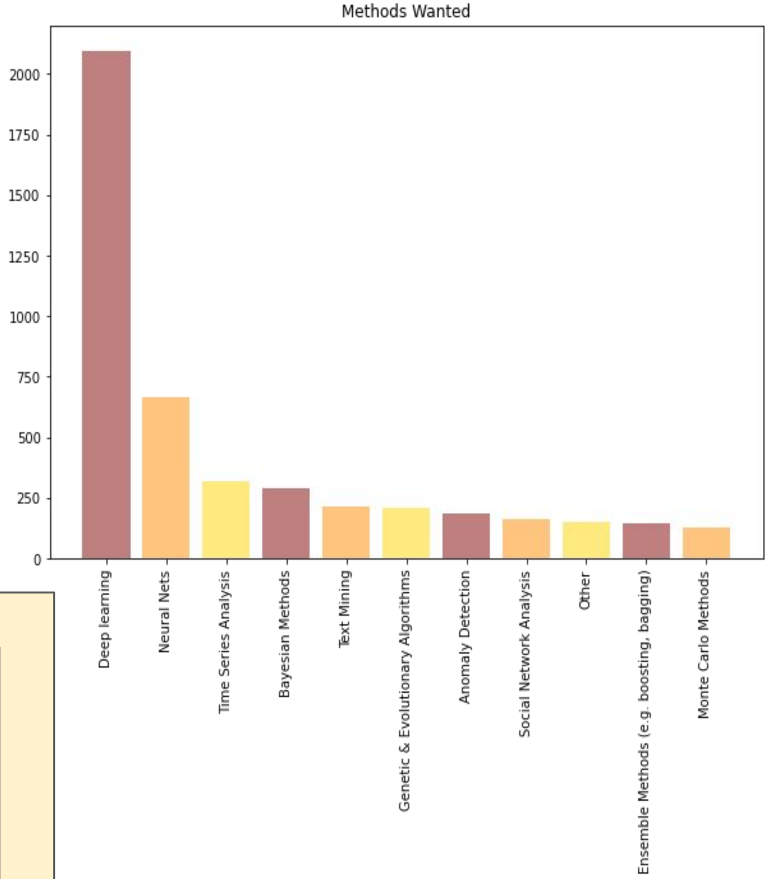
Most common age group among Kagglers is 30-40 and the second most common is 18-30, which confirms that most of Kaggle users are **young adults/adults**.





According to this graph, IMLearning should mostly target its advertisements towards data scientists of **higher education** to maximize the number of potential customers.



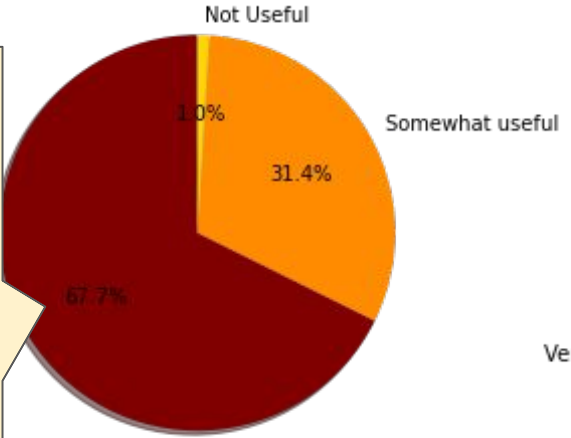


The dataset provided us with which tools and methods kagglers would like to learn within the next year. These would be useful in decided what content to provide on platform being made. There is a prominent interest in Tensorflow and Deep Learning

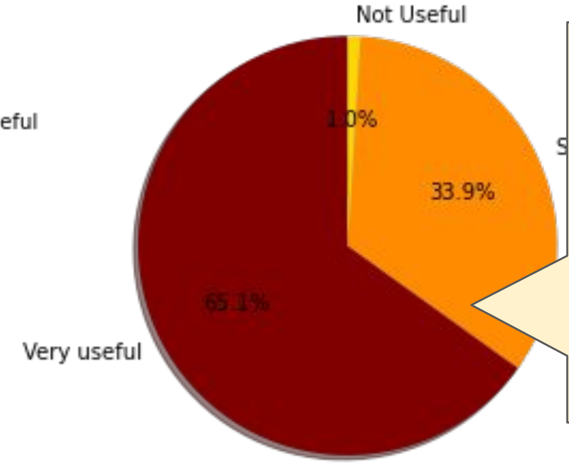
What platforms are the best?

- The data collected from Kagglers also have provided some useful insights about their opinions on various learning platforms that provide similar services as IMLearning.
- Such insights would be clearly beneficial

Online course format is among the most useful platform types due to its similarity to remote work.

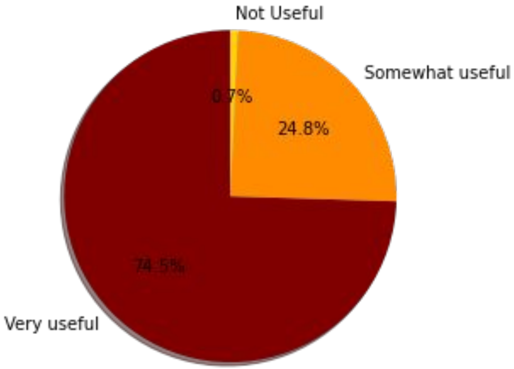


Courses: 2941 entries



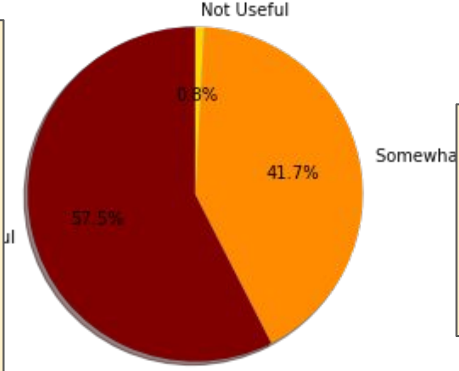
Including community section where users can ask and answers questions increases usefulness.

Stack overflow: 2996 entries



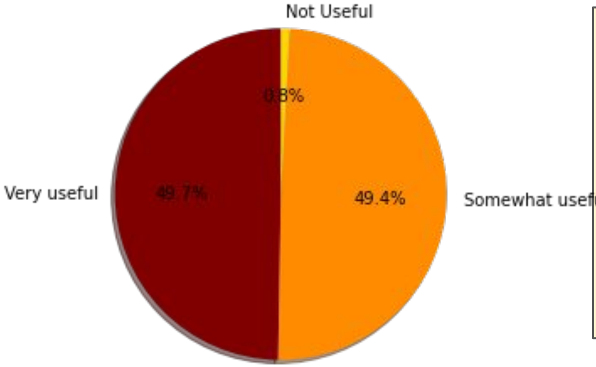
Personal Projects: 2499 entries

Surprisingly enough, Kaggle, where all of the subjects in this dataset was registered, is not considered the most useful learning platform according to the results



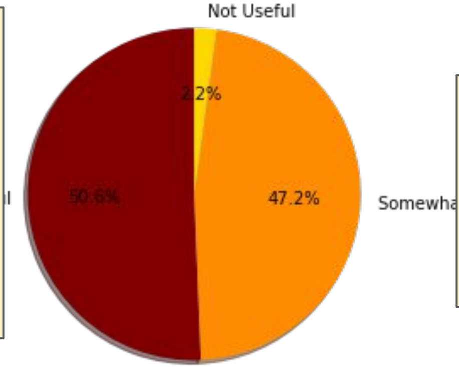
Kaggle: 3168 entries

Platform could encourage exploration of personal projects



Blogs usefulness: 2531 entries

Blogs and Youtube as a data science learning platform appears to have received mixed reception from Kagglers.



Youtube: 2413 entries

Ability to create own content like on YouTube is not necessary

Preprocessing

- After a thorough analysis of the datatypes of the given features and what kind of values they have, 3 types of conversions were done to the categorical data. First, the ordinal features were given values starting from 0 to (number-of-values - 1).
- With some of the categorical non-ordinal features, one hot encoder was used after filling the missing values with mode, and with the rest we used the encoder provided with **sklearn.OrdinalEncoder** as it increased our accuracy compared to OneHotEncoding everything.
- At this point, all of the categorical data was converted to numerical so **KNN imputer** was used to fill in the remaining missing values. Using KNN gave better accuracy as opposed to imputing with mean, mode, or median.
- Data was then normalized.

Feature Selection and Extraction

- New features were added to features like “MLTechniqueSelect”, which counts the number of ML techniques used by counting the number of commas in each entry and adding 1. Four other multi-selection features were treated similarly.
- Through our work with dataset and checking for correlation, we found the most important features are: TitleFit, RemoteWork, WorkProductionFrequency, Number of Algor, Number of techniques, CurrentEmployerType_Employed by a company that doesn't perform analystics (negative correlation)
- PCA was attempted for feature extraction but did not yield good results.
- ID was dropped for its irrelevance and CodeWriterwas dropped due to all the values being the same

Our Algorithm

- Multiple machine learning models were tried including KNN, Random Forest, PCA with KNN, however we achieved our best results using voting with multiple regressors. Our voting regressor used **GradientBoosting, RandomForest, LinearRegressor, and KNN**.
- The cross validation results are as follows, corresponding to best submission on kaggle, with negative RMSE:

```
[[-1.88053647 -1.99315499 -2.05188144 -1.98489421 -1.89119021 -1.98276046
-2.02403185 -2.04615849 -2.01588837 -2.13458057]
```

Average: -2.0005

Analysis of System Performance

- The use of multiple regressors and voting proved to be a good choice in our system, giving us a significantly better accuracy compared to single regressors.
- Additionally, the new features we added reduced the error and became one of most correlated features with JobSatisfaction.
- That being said, the system could be improved upon, mostly in the feature handling stage, different methods for feature extraction could have been experimented with.