# Kaggle ML and Data Science Survey, 2017

### Context

For the first time, Kaggle conducted an industry-wide survey to establish a comprehensive view of the state of data science and machine learning. The survey received over 16,000 responses and we learned a ton about who is working with data, what’s happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field.

To share some of the initial insights from the survey, we’ve worked with the folks from [The Pudding](https://pudding.cool/) to put together this interactive report. They’ve shared all of the kernels used in the report here.

### Content

The data includes 5 files:

* schema.csv: a CSV file with survey schema. This schema includes the questions that correspond to each column name in both the multipleChoiceResponses.csv and freeformResponses.csv.
* multipleChoiceResponses.csv: Respondents' answers to multiple choice and ranking questions. These are non-randomized and thus a single row does correspond to all of a single user's answers. -freeformResponses.csv: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.
* conversionRates.csv: Currency conversion rates (to USD) as accessed from the R package "quantmod" on September 14, 2017
* RespondentTypeREADME.txt: This is a schema for decoding the responses in the "Asked" column of the schema.csv file.

### Kernel Awards in November

In the month of November, we’re awarding $1000 a week for code and analyses shared on this dataset via [Kaggle Kernels](https://www.kaggle.com/kaggle/kaggle-survey-2017/kernels). Read more about this month’s [Kaggle Kernels Awards](https://www.kaggle.com/about/datasets-awards/kernels) and help us advance the state of machine learning and data science by exploring this one of a kind dataset.

### Methodology

* This survey received 16,716 usable respondents from 171 countries and territories. If a country or territory received less than 50 respondents, we grouped them into a group named “Other” for anonymity.
* We excluded respondents who were flagged by our survey system as “Spam” or who did not answer the question regarding their employment status (this question was the first required question, so not answering it indicates that the respondent did not proceed past the 5th question in our survey).
* Most of our respondents were found primarily through Kaggle channels, like our email list, discussion forums and social media channels.
* The survey was live from August 7th to August 25th. The median response time for those who participated in the survey was 16.4 minutes. We allowed respondents to complete the survey at any time during that window.
* We received salary data by first asking respondents for their day-to-day currency, and then asking them to write in either their total compensation.
  + We’ve provided a csv with an exchange rate to USD for you to calculate the salary in US dollars on your own.
  + The question was optional
* Not every question was shown to every respondent. In an attempt to ask relevant questions to each respondent, we generally asked work related questions to employed data scientists and learning related questions to students. There is a column in the schema.csvfile called "Asked" that describes who saw each question. You can learn more about the different segments we used in the schema.csv file and RespondentTypeREADME.txt in the data tab.
* To protect the respondents’ identity, the answers to multiple choice questions have been separated into a separate data file from the open-ended responses. We do not provide a key to match up the multiple choice and free form responses. Further, the free form responses have been randomized column-wise such that the responses that appear on the same row did not necessarily come from the same survey-taker.

# Novice to Grandmaster- What Data Scientists say?

Kaggle is the world's largest Data Science platform with more than 1 million users, and it is an excellent platform for students like me to learn and grow in the field of Data Science and Machine Learning. It has users from various domains,like statisticians,Data Scientists and Machine Learning Practitioners.This dataset published by Kaggle is a gem for people like me, who like to analyse and investigate data. In this notebook, we will try to find some trending or some common questions, each budding data scientist would like to know, like the most used tools, the resources to learn data science ,etc.

The biggest problem that we might face is fake and bogus responses. As it is a survey, not everyone will answer with proper credentials, and thus I assume that there will be a lot many outlier. Let's dive in straight into the pool of data and gain some insights..

### Who are Data Scientists?[¶](https://www.kaggleusercontent.com/kf/2728011/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..MBczGf37xQMhFILVef_CCg.tbhZiR3wDxSrDMBFRifTVxNcduRdKTCpe9pyC-sPBK0qDUjAd_PU0gGdTEgKC61UvAnXf1iasiklxeWUg-3IAlFXy_zUFObGeI7DkUV_UPMnzdQQn6ooNXqVvlhDwsWWLJRIq5g6-aBMVjyemhGZHQ.VQHYXvro42hr2zoA7Bgf-g/__results__.html" \l "Who-are-Data-Scientists?" \t "https://www.kaggleusercontent.com/kf/2728011/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..MBczGf37xQMhFILVef_CCg.tbhZiR3wDxSrDMBFRifTVxNcduRdKTCpe9pyC-sPBK0qDUjAd_PU0gGdTEgKC61UvAnXf1iasiklxeWUg-3IAlFXy_zUFObGeI7DkUV_UPMnzdQQn6ooNXqVvlhDwsWWLJRIq5g6-aBMVjyemhGZHQ.VQHYXvro42hr2zoA7Bgf-g/_self)

A data scientist is a statistician or a programmer, who cleans, manages and organizes data, perform descriptive statistics and analysis to develop insights,build predictive models and solve business related problems. Let's see what do Data Scientists on kaggle say..