





12 active competitions		Sort by	Prize
Active	All	Entered	Unlaunched
All Categories		Search	Q
	<b>Passenger Screening Algorithm Challenge</b> Improve the accuracy of the Department of Homeland Security's threat recognition algorithms Featured · 5 months to go	<b>\$1,500,000</b> 89 teams	
	<b>Zillow Prize: Zillow's Home Value Prediction (Zestimate)</b> Can you improve the algorithm that changed the world of real estate? Featured · 6 months to go	<b>\$1,200,000</b> 1,425 teams	
	<b>Planet: Understanding the Amazon from Space</b> Use satellite data to track the human footprint in the Amazon rainforest Featured · 9 days to go	<b>\$60,000</b> 840 teams	
	<b>Instacart Market Basket Analysis</b> Which products will an Instacart consumer purchase again? Featured · a month to go	<b>\$25,000</b> 1,307 teams	

# The Beginner's Guide to Kaggle


[elitedatascience.com](https://elitedatascience.com)




6 mins read

Kaggle, a popular platform for [data science competitions](#), can be intimidating for beginners to get into.

After all, some of the listed competitions have over \$1,000,000 prize pools and hundreds of competitors.

Top teams boast decades of combined experience, tackling ambitious problems such as improving airport security or analyzing satellite data.

12 active competitions		Sort by	Prize
Active	All	Entered	Unlaunched
All Categories		Search	Q
	<b>Passenger Screening Algorithm Challenge</b> Improve the accuracy of the Department of Homeland Security's threat recognition algorithms Featured · 5 months to go	<b>\$1,500,000</b> 89 teams	

	<b>Zillow Prize: Zillow's Home Value Prediction (Zestimate)</b> Can you improve the algorithm that changed the world of real estate? <small>Featured · 6 months to go</small>	<b>\$1,200,000</b> 1,425 teams
	<b>Planet: Understanding the Amazon from Space</b> Use satellite data to track the human footprint in the Amazon rainforest <small>Featured · 9 days to go</small>	<b>\$60,000</b> 840 teams
	<b>Instacart Market Basket Analysis</b> Which products will an Instacart consumer purchase again? <small>Featured · a month to go</small>	<b>\$25,000</b> 1,307 teams

It's no surprise that some beginners hesitate to get started on Kaggle. They have reasonable concerns such as:

- How do I even start?
- Will I be up against teams of experienced Ph.D researchers?
- Is it worth competing if I don't have a realistic chance of winning?
- Is this what data science is all about? (If I don't do well on Kaggle, do I have future in data science?)
- How can I improve my rank in the future?

Well, if you've ever had any of those questions, you're in the right place.

In this guide, we'll break down everything you need to know about getting started, improving your skills, and enjoying your time on Kaggle.

## Kaggle vs. "Typical" Data Science

First, we need to make something very clear:

*Kaggle competitions have important differences from "typical" data science, but they still provide valuable*

*experience if you approach them with the right mindset.*

Let us explain:

### ***Kaggle competitions***

By nature, competitions (with prize pools) must meet several criteria.

- **Problems must be difficult.** Competitions shouldn't be solvable in a single afternoon. To get the best return on investment, host companies will submit their biggest, hairiest problems.
- **Solutions must be new.** To win the latest competitions, you'll usually need to perform extended research, customize algorithms, train advanced models, etc.
- **Performance must be relative.** Competitions must crown a winner, so your solution will be scored against others'.

### ***"Typical" data science***

In contrast, day-to-day data science doesn't need to meet those same criteria.


- **Problems *can* be easy.** In fact, data scientists should try to identify low-hanging fruit: impactful projects that can be solved quickly.
- **Solutions *can* be mature.** Most common tasks (e.g. exploratory analysis, data cleaning, A/B testing, classic algorithms) already have proven frameworks. There's need to reinvent the wheel.
- **Performance *can* be absolute.** A solution can be very valuable even if it simply beats a previous benchmark.

Kaggle competitions encourage you to squeeze out every last drop of performance, while typical data science encourages efficiency and maximizing business impact.

### ***So is Kaggle worth it?***

Despite the differences between Kaggle and typical data science, Kaggle can still be a great learning tool for beginners.

- **Each competition is self-contained.** You don't need to scope your own project and collect data, which frees you up to focus on other skills.
- **Practice is practice.** The best way to learn data science is to learn by doing. As long as you don't stress out about winning every competition, you can still practice interesting problems.
- **The discussions and winner interviews are enlightening.** Each competition has its own discussion board and [debriefs with the winners](#). You can peek into the thought-processes of more experienced data scientists.

<div data-bbox="755 1165 787 1207">1</div> <div data-bbox="402 1186 701 1354"><b>The Nature Conservancy Fisheries Monitoring Competition, 1st Place Winner's Interview: Team 'Towards Robust-Optimal Learning of Learning'</b></div> <div data-bbox="441 1358 659 1381">Kaggle Team   07.07.2017</div> <div data-bbox="371 1392 719 1682"><p>This year, The Nature Conservancy Fisheries Monitoring competition challenged the Kaggle community to develop algorithms that automatically detects and classifies species of sea life that fishing boats catch. Illegal and unreported fishing practices threaten marine ecosystems. These algorithms would help increase The Nature Conservancy's capacity to analyze data from camera-based monitoring systems. In this winners' interview, first place team, 'Towards Robust-Optimal Learning of Learning' (Gediminas Pekšys, Ignas Namajūnas, Jonas Bialopetavičius), shares details of their approach like how they needed to have a ...</p></div>	<div data-bbox="1291 1165 1323 1207">3</div> <div data-bbox="902 1186 1252 1325"><b>2017 Data Science Bowl, Predicting Lung Cancer: 2nd Place Solution Write-up, Daniel Hammack and Julian de Wit</b></div> <div data-bbox="967 1331 1185 1354">Kaggle Team   06.29.2017</div> <div data-bbox="862 1375 1289 1518"></div>
	<div data-bbox="1291 1606 1323 1648">5</div> <div data-bbox="902 1623 1252 1732"><b>Stacking Made Easy: An Introduction to StackNet by Competitions Grandmaster Marios Michailidis (KazAnova)</b></div>

Kaggle winner interviews

# How to Get Started on Kaggle

Next, we'll give you a step-by-step action plan for gently ramping up and competing on Kaggle.

## ***Step 1: Pick a programming language.***

First, we recommend picking one programming language and sticking with it. Both Python and R are popular on Kaggle and in the broader data science community.

If you're starting with a blank slate, we recommend **Python** because it's a general-purpose programming language that you can use from end-to-end.

## ***Step 2: Learn the basics of exploring data.***

The ability to load, navigate, and plot your data (i.e. exploratory analysis) is the first step in data science because it informs the various decisions you'll make throughout model training.

If you go the route of Python, then we recommend the [Seaborn](#) library, which was designed specifically for this purpose. It has high-level functions for plotting many of the most common and useful charts.

## ***Step 3: Train your first machine learning model.***

Before jumping into Kaggle, we recommend training a model on an easier, more manageable dataset. This will allow you to become familiar with machine learning libraries and the lay of the land.

The key is to start developing good habits, such as splitting your dataset into separate training and testing sets, cross-validating to avoid overfitting, and using proper performance metrics.

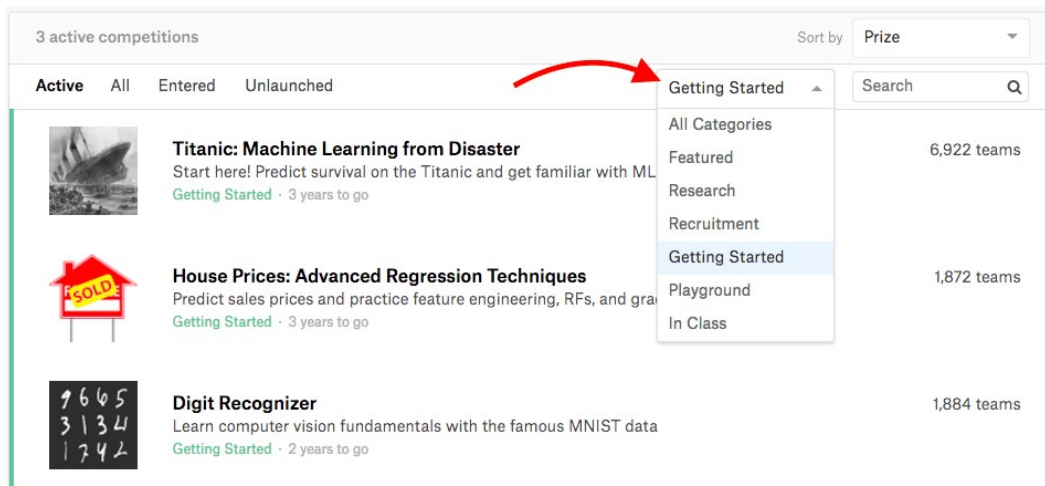
For Python, the best general-purpose machine learning library is [Scikit-Learn](#).

## ***Step 4: Tackle the 'Getting Started' competitions.***

Now we're ready to try Kaggle competitions, which fall into several categories. The most common ones are:

- **Featured** - These are usually sponsored by companies, organizations, or even governments. They have the largest prize pools.
- **Research** - These are research-oriented and have little to no prize money. They also have non-traditional submission processes.
- **Recruitment** - These are sponsored by companies who want to hire data scientists. These are still relatively uncommon.
- **Getting Started** - These are structured like featured competitions, but they have no prize pools. They feature easier datasets, plenty of tutorials, and rolling submission windows so you can enter them at any time.

The 'Getting Started' competitions are great for beginners because they give you a low-stakes environment to learn, and they are also supported by many [community-created tutorials](#).



The screenshot shows the Kaggle competitions interface. At the top, it says '3 active competitions'. Below this are tabs for 'Active', 'All', 'Entered', and 'Unlaunched'. A red arrow points to a dropdown menu that is open, showing the following options: 'Getting Started' (selected), 'All Categories', 'Featured', 'Research', 'Recruitment', 'Playground', and 'In Class'. The main list of competitions includes:

Competition	Teams
<b>Titanic: Machine Learning from Disaster</b> Start here! Predict survival on the Titanic and get familiar with ML <i>Getting Started</i> · 3 years to go	6,922 teams
<b>House Prices: Advanced Regression Techniques</b> Predict sales prices and practice feature engineering, RFs, and gradient boosting <i>Getting Started</i> · 3 years to go	1,872 teams
<b>Digit Recognizer</b> Learn computer vision fundamentals with the famous MNIST data <i>Getting Started</i> · 2 years to go	1,884 teams

### ***Step 5: Compete to maximize learnings, not earnings.***

With that foundation laid, it's time to progress to 'Featured' competitions. In general, these will require much more time and effort to rank well.

For that reason, we recommend picking your battles wisely. Enter competitions that will expose you to techniques and technologies that align with your long-term goals.

While prize money is nice, the more valuable (and reliable) reward will be the skills you'll develop for your career.

## **Tips for Enjoying Kaggle**

Finally, we'll cover our 7 favorite tips for making the most out of your time on Kaggle.

### ***Tip #1: Set incremental goals.***

If you've ever played an addicting video game, you'll know the power of incremental goals. That's how great games get you hooked. Each goal is big enough for a sense of accomplishment, yet realistic enough to be within reach.





Most Kaggle participants will never win a single competition, and that's completely fine. If you set that as your very first milestone, you may feel discouraged and lose motivation after a few tries.

Incremental targets make the journey more enjoyable. For example:

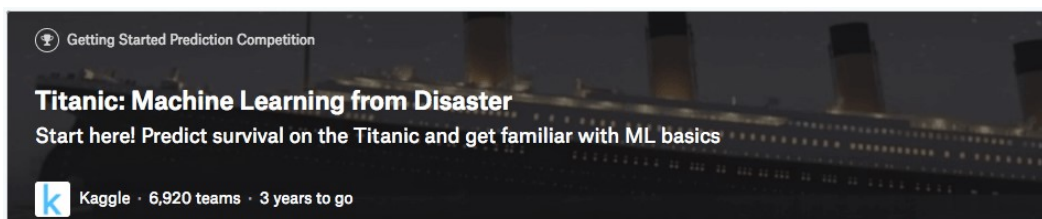
- Make a submission that beats the benchmark solution.
- Score in the top 50% in one competition.
- Score in the top 25% in one competition.
- Score in the top 25% in three competitions.
- Score in the top 10% in one competition.
- Win a competition!

This strategy will allow you to measure your progress and improvement along the way.

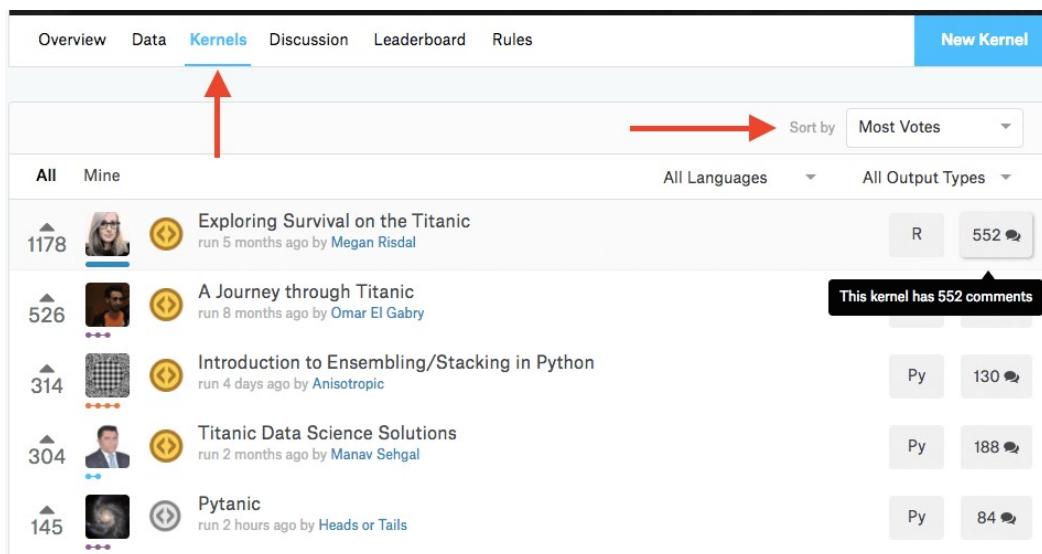
***Tip #2: Review most voted kernels.***

Kaggle has a cool feature in which participants can submit "kernels," which are short scripts that explore a concept, showcase a technique, or even share a solution.

When you start a competition or when you hit a plateau, reviewing popular kernels can spark more ideas.







### ***Tip #3: Ask questions on the forums.***

Don't be afraid to ask "stupid" questions.

After all, what's the worst thing that could happen? Maybe you get ignored... and that's all.

On the other hand, you have plenty to gain, including advice and coaching from more experienced data scientists.

### ***Tip #4: Work solo to develop core skills.***

In the beginning, we recommend working alone. This will force you to tackle every step of the applied machine learning process, including exploratory analysis, data cleaning, feature engineering, and model training.

If you start teaming up too early, you could miss opportunities to develop those cornerstone skills.

### ***Tip #5: Team up to push your boundaries.***

With that said, teaming up in future competitions can be a great way to push your boundaries and learn from others. Many past winners have been teams who joined forces to combine their knowledge.

In addition, once you master the technical skills of machine learning, you can collaborate with others who may have more domain knowledge than you do, further expanding your opportunities.

***Tip #6: Remember that Kaggle can be a stepping stone.***

Remember, you're not necessarily committing to be a long-term Kagglers. If you find out that you dislike the format, then it's no big deal.

In fact, many people use Kaggle as a stepping stone before moving onto their own projects or becoming full-time data scientists.

This is another reason to focus on learning as much as you can. For the long run, it's better to target competitions that will give you relevant experience than to chase the biggest prize pools.

***Tip #7: Don't worry about low ranks.***

Some beginners never start because they're worried about low ranks showing up in their profile. Of course, competition anxiety is a real phenomenon, and it isn't limited to Kaggle.

However, low ranks are really not a big deal. No one else will judge you because they were all beginners at one point.



Even so, if you're still really worried about low rankings in your profile, you could also create a separate practice account for learning

the ropes. Once you feel comfortable, you can start using your "main account" to build your trophy case.

(Again, this is totally unnecessary!)

## Conclusion

In this guide, we shared the 5 steps for getting started on Kaggle:

1. Pick a programming language.
2. Learn the basics of exploring data.
3. Train your first machine learning model.
4. Tackle the 'Getting Started' competitions.
5. Compete to maximize learnings, not earnings.

Finally, we shared our 7 favorite tips for enjoying your time on the platform:

- Set incremental goals.
- Review most voted kernels.
- Asks questions on the forums.
- Work solo to develop core skills.
- Team up to push your boundaries.
- Remember that Kaggle can be a stepping stone.
- Don't worry about low ranks.

