



# Quick start guide to Kaggle – Titanic

BY KUNAAL NAIK

Lifeaholic Channel



LinkedIn profile of Kunaal Naik. The profile picture shows a man with glasses and a dark suit. The background of the profile is an abstract painting with vibrant colors like red, yellow, green, and black. The profile name is Kunaal Naik, and the headline is Analytics Practitioner, Lifeaholic Evangelist, Learner, YouTuber and Apprentice Philosopher. The location is Brillio • Institute of Aeronautical Engineering, Bengaluru, Karnataka, India • 500+ connections.



fxexcel@gmail.com

The humble me

## Career Square

- Current job
- Fun X Excel
- Jigsaw
- Work Life Balance
- Others

## Personal Branding

- LinkedIn
- Writing
- Raise your standards
- Social Media
- Experiments

## Passion (Why?)


## Family and Friends

- Wife/Girlfriend
- Relationships
- Office colleges
- Jigsaw Community
- Health and Spirituality

## Learning

- Finance
- YouTube, Lynda – Calendarized
  - Reading and implementing
  - Musical Instrument
  - Travel

# Lifeaholic



“My attempt will be to make  
you fall in love with Analytics.

”

KUNAAL NAIK


**Guiding Principles:**

1. Does not cover statistics in details. Rather focuses on the problem solving approach and submission to Kaggle.
2. Primary tool – R and Excel. SAS and Python codes also on GitHub.
3. Work along session. If you miss out just start from the next topic.



# Guidelines


- ▶ Follow along session
- ▶ R for building out models and Excel for Data Manipulation
- ▶ We will make around 10 submissions in this session
- ▶ We explore how to run various models and compare which performs best
- ▶ We will do some amount of Feature Engineering
- ▶ All files are placed on GitHub. In case you miss any step you can download the code and input files.



“We have a lot to cover. Lets keep time consuming doubts towards the end of the session.”

KUNAAL

The session is getting recorded and there are many participants. I want to ensure we all make our submission on Kaggle. 😊



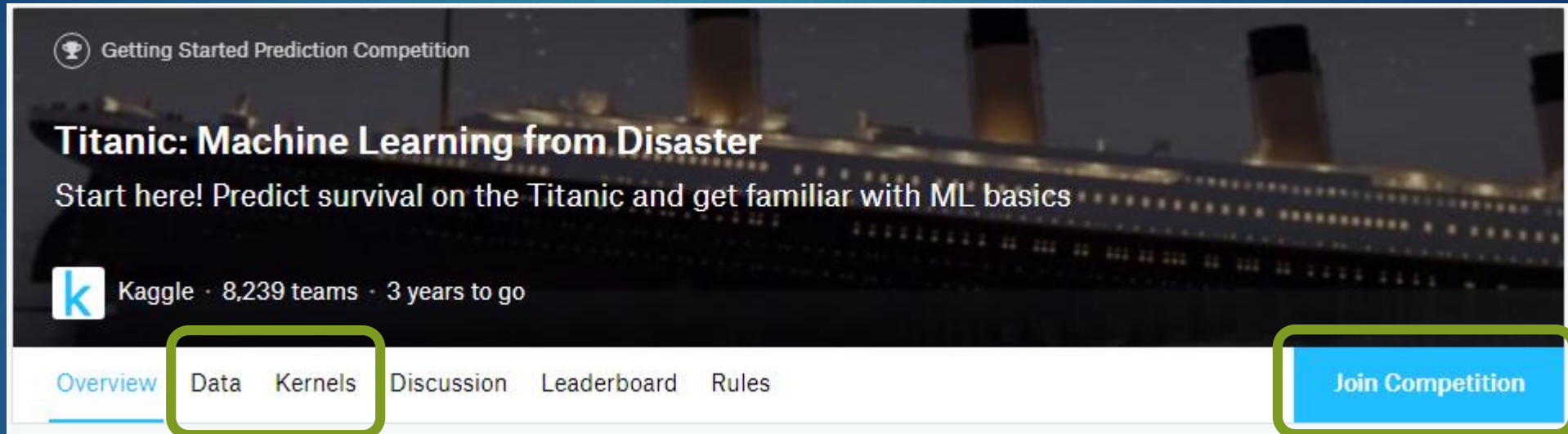
# Kaggle and how to use it smartly

Unless you get exposed to a variety of datasets for Analytics, you will not become confident during your interviews.

**What to do on Kaggle?**

# Select a competition – E.g.. Titanic

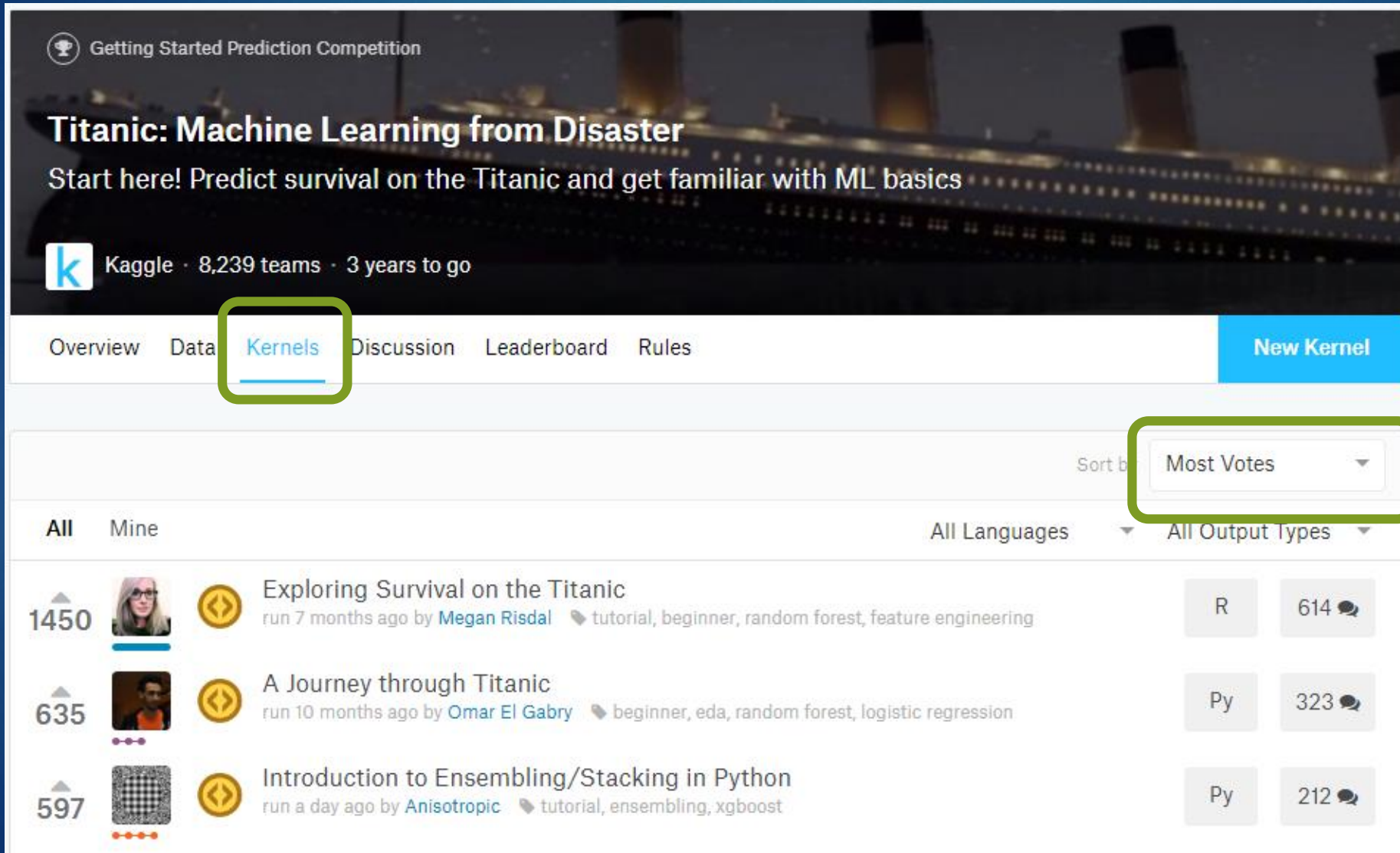
## | Read the overview



- Join completion
- Choose your tool
- Download data and get started
- You have an option to run the codes on Kaggle itself
- Take help from Kernel which are written by participants for others to learn and use
- Try and be at least on top 25 in the leaderboard



# How to start on Kaggle



The screenshot shows the Kaggle interface for the 'Titanic: Machine Learning from Disaster' competition. The 'Kernels' tab is selected and highlighted with a green box. Below the navigation bar, a dropdown menu for sorting is also highlighted with a green box, showing 'Most Votes' as the selected option. The list of kernels is displayed below, with the top three kernels shown. Each kernel entry includes a rank, a profile picture, a kernel title, a description, and the number of votes.

Getting Started Prediction Competition

## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics




Kaggle · 8,239 teams · 3 years to go

Overview Data **Kernels** Discussion Leaderboard Rules

New Kernel

Sort by Most Votes

All Mine All Languages All Output Types

Rank	Profile	Kernel Title	Description	Language	Votes
1450		Exploring Survival on the Titanic	run 7 months ago by Megan Risdal · tutorial, beginner, random forest, feature engineering	R	614
635		A Journey through Titanic	run 10 months ago by Omar El Gabry · beginner, eda, random forest, logistic regression	Py	323
597		Introduction to Ensembling/Stacking in Python	run a day ago by Anisotropic · tutorial, ensembling, xgboost	Py	212

- Go to Kernels
- Select the Most Votes(in Sort by)
- Choose your preferred Tool (R, Python)
- Select the kernel

# How to start on Kaggle

**Megan Risdal**  
**Exploring Survival on the Titanic**  
last run 7 months ago · R notebook · 262544 views  
using data from [Titanic: Machine Learning from Disaster](#) · Public

1450 voters

[Report](#) [Code](#) [Data \(1\)](#) [Output \(1\)](#) [Comments \(614\)](#) [Log](#) [Versions \(7\)](#) [Forks \(2743\)](#) [Fork Script](#)

Tags: [tutorial](#) [feature engineering](#) [beginner](#) [random forest](#)

Report

## Exploring the Titanic Dataset

*Megan L. Risdal*

6 March 2016

- 1 Introduction
  - 1.1 Load and check data
- 2 Feature Engineering
  - 2.1 What's in a name?
  - 2.2 Do families sink or swim together?
  - 2.3 Treat a few more variables ...

- You find a variety of sections
- Each section will have components of the analysis that is required
- When you are working in a real environment, the practice given here will give you good clarity
- **Fork Script** – Gives you an option to copy the script and use it for your submissions

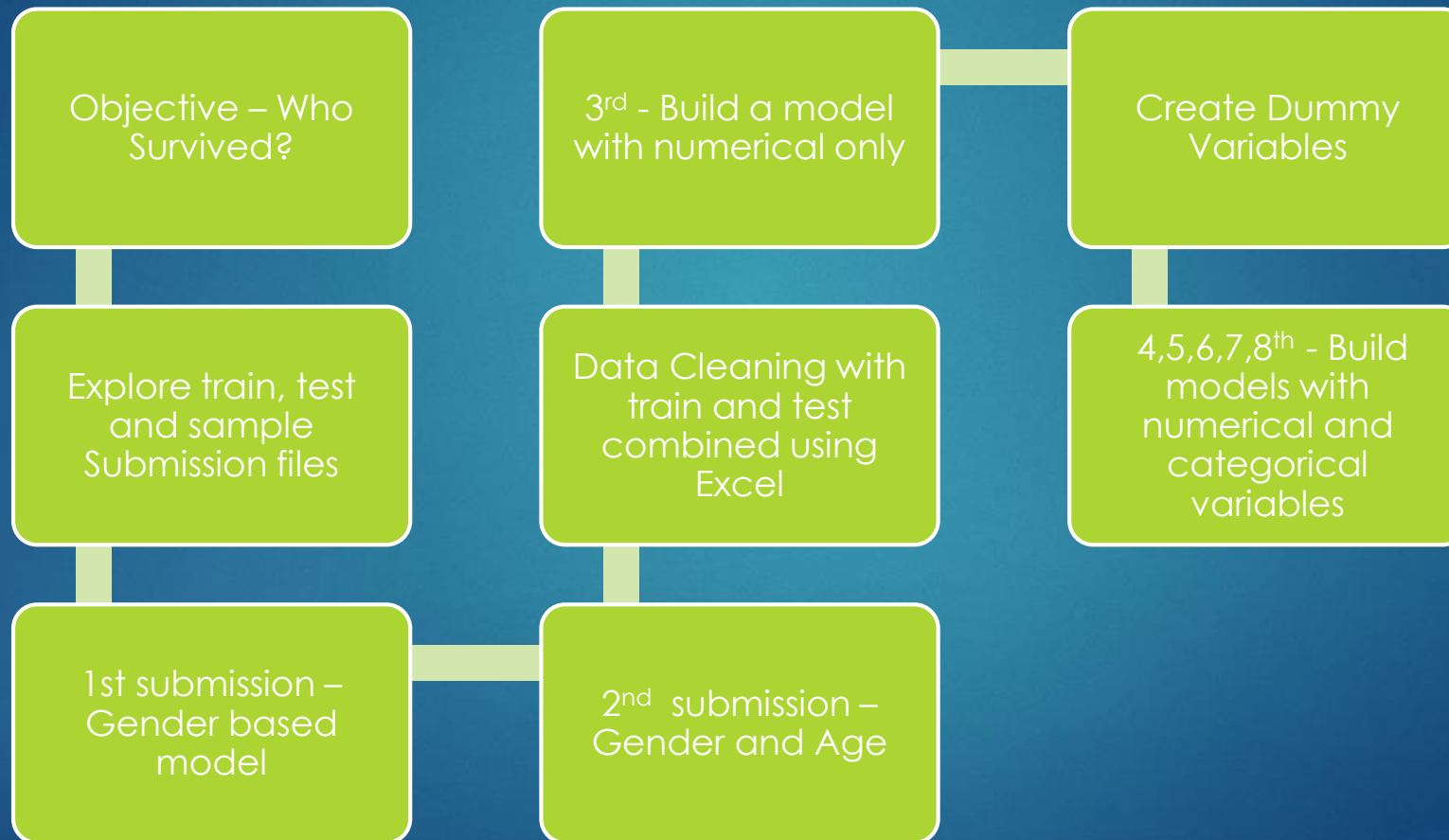
**Call to action:** Create a fork, run as in and make a submission




# Process for solving Kaggle problems and making submissions

EVERYTHING WORKS LIKE A CHARM IF YOU HAVE A PROCESS!

# Warming up - Build the first model as quick as possible





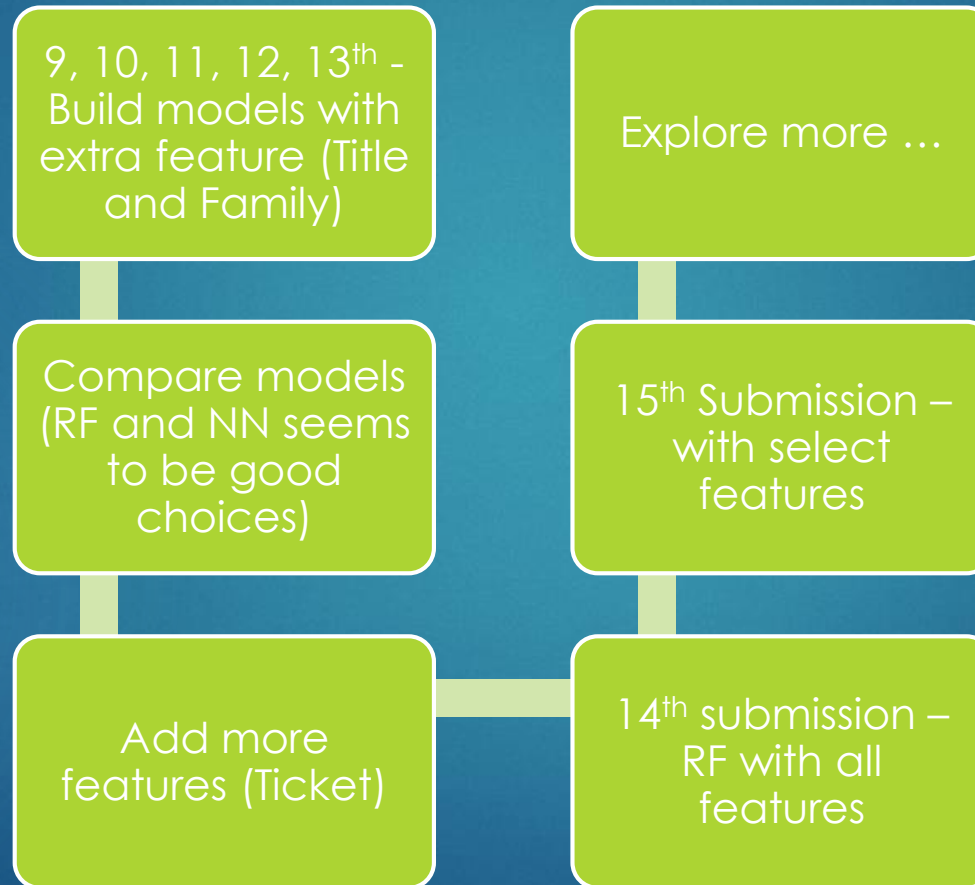


“Today you will learn about the importance of Excel as a Logical Thinking Tool.”

KUNAAL

Often we underestimate what Excel can do. I feel we just know enough options because it is such an ocean.

# Getting serious – Compare models and introduce feature engineering



# What to explore?

## More Feature Engineering

- Surname to find families
- Mother variable
- Categories Age and Fare

## Advanced Models

- Deep Machine Learning – H2O, keras package on R


## Using Statistics to impute missing values

- Use regression to fill the missing values for Age

# References

- ▶ Logistic Regression : <https://www.kaggle.com/iliasemenov/logistic-regression>
- ▶ Caret Package : <http://topepo.github.io/caret/available-models.html>
- ▶ Feature Selection : <https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/>
- ▶ Performance Plots : <https://machinelearningmastery.com/compare-models-and-select-the-best-using-the-caret-r-package/>
- ▶ Score 0.81 : <https://ahmedbesbes.com/how-to-score-08134-in-titanic-kaggle-challenge.html>
- ▶ Score 0.82 : <https://www.kaggle.com/pliptor/divide-and-conquer-0-82296/code>





“Hope you solve many more  
Kaggle problems!

”

KUNAAL NAIK

**Remember:**

1. Choose a tool(R/Python) – They are going to be here for next 10 years.
2. Use Kernels. Find the those that have a high score and understand the methods that worked.
3. Use Train/Cross Validation before you submit to Kaggle. In real life you wont have a scoring engine like Kaggle to track your score.

Lifeaholic Channel







...



**Kunaal Naik**

Analytics Practitioner, Lifeaholic Evangelist, Learner, YouTuber and Apprentice Philosopher

Brillio • Institute of Aeronautical Engineering  
Bengaluru, Karnataka, India • 500+ 



[fxexcel@gmail.com](mailto:fxexcel@gmail.com)

Stay in touch!