Data Scientist - Take Home Challenge

## 

## Instruction

Hello Candidate!

Here are some fun and intellectually interesting questions related to a singular dataset. Since we are the company behind Pokemon Go, you are probably not surprised that the dataset is on Pokemon. There are four questions and you are expected to spend no more than two to three hours on this challenge. Please submit your write-up to your recruiter within 48 hours of receipt.

A few things to note:

* The dataset is a single csv file attached with this take-home challenge. If you have difficulty opening this csv file or have any questions during the take-home challenge, please reach out to the recruiter.
* Tools for analysis: unless specified in the questions, feel free to use whatever analysis software you think makes the most sense.
* Prior knowledge: Having prior knowledge on this topic can be helpful but is not necessary. If you have applicable contextual information, feel free to mention it in your write-up. However, note that we are more interested in the thoroughness of your data analysis skills here.
* Outside help: internet research for generic questions is fine but you are expected to complete this challenge without consulting other individuals for help or collaboration.
* Submission and presentation: The only submission you will need to send to your recruiter is your completed answers to the four questions below. The submission can be in any format you prefer (a document, a presentation, a dashboard, etc) and can be several files if you wish- whichever can convey your messages clearly and effectively. If you excel in this take-home challenge, you will have the chance to present your findings to your interviewer in your next interview. Treat that interview as a two-way conversation: you will start with a few minutes of presentation, and then your interviewer may ask you a few follow-up questions and you will respond based on your understanding of the data. You will be evaluated on the quality of your data analysis as well as the clarity and depth of your discussion. Please note, the discussion on the take-home challenge will only be a portion of the next interview - please be prepared to answer a few additional technical questions.

## Dataset

The Pokémon Dataset is focused on the stats and features of the Pokémon in the Pokémon RPG games until Generation 6.

This database includes 21 variables for each of the 721 Pokémon of the first six generations, plus the Pokémon ID and its name. These variables are briefly described here:

* Number. Pokémon ID in the Pokédex.
* Name. Name of the Pokémon.
* Type\_1. Primary type.
* Type\_2. Second type, in case the Pokémon has it.
* Total. Sum of all the base stats (Health Points, Attack, Defense, Special Attack, Special Defense, and Speed).
* HP. Base Health Points.
* Attack. Base Attack.
* Defense. Base Defense.
* Sp\_Atk. Base Special Attack.
* Sp\_Def. Base Special Defense.
* Speed. Base Speed.
* Generation. Number of the generation when the Pokémon was introduced.
* isLegendary. Boolean that indicates whether the Pokémon is Legendary or not.
* Color. Color of the Pokémon according to the Pokédex.
* hasGender. Boolean that indicates if the Pokémon can be classified as female or male.
* Pr\_male. In case the Pokémon has Gender, the probability of its being male. The probability of being female is, of course, 1 minus this value.
* Egg\_Group\_1. Egg Group of the Pokémon.
* Egg\_Group\_2. Second Egg Group of the Pokémon, in case it has two.
* hasMegaEvolution. Boolean that indicates whether the Pokémon is able to Mega-evolve or not.
* Height\_m. Height of the Pokémon, in meters.
* Weight\_kg. Weight of the Pokémon, in kilograms.
* Catch\_Rate. Catch Rate.
* Body\_Style. Body Style of the Pokémon according to the Pokédex.

## Question 1

Suppose this data is a SQL table called ‘PokemonStats’. In an SQL dialect you are most comfortable with, find the top 3 Pokemon in terms of total stats of each type (primary type, Type\_1). Your answer should include: 1) the SQL dialect you are using; 2) The SQL query used to answer this question; 3) The returned result.

1. The dialect I used is MySQL
2. SQL query:

|  |
| --- |
| SELECT p.Type\_1, p.Name, p.Total  FROM PokemonStats as p  WHERE(SELECT count(\*) FROM PokemonStats as k  WHERE k.Type\_1 = p.Type\_1 and k.Total >= p.Total) <= 3  ORDER BY p.Type\_1 ASC, p.Total DESC |

1. The returned result

|  |
| --- |
| Type\_1 Name Total  0 Bug Genesect 600  1 Bug Volcarona 550  2 Bug Yanmega 515  3 Dark Yveltal 680  4 Dark Darkrai 600  5 Dark Hydreigon 600  6 Dragon Rayquaza 680  7 Dragon Reshiram 680  8 Dragon Zekrom 680  9 Electric Zapdos 580  10 Electric Raikou 580  11 Electric Thundurus 580  12 Fairy Xerneas 680  13 Fairy Florges 552  14 Fairy Togekiss 545  15 Fighting Lucario 525  16 Fighting Mienshao 510  17 Fire Ho-Oh 680  18 Fire Heatran 600  19 Fire Volcanion 600  20 Flying Tornadus 580  21 Flying Noivern 535  22 Flying Noibat 245  23 Ghost Giratina 680  24 Ghost Dusknoir 525  25 Ghost Chandelure 520  26 Grass Shaymin 600  27 Grass Virizion 580  28 Grass Tangrowth 535  29 Ground Groudon 670  30 Ground Landorus 600  31 Ground Rhyperior 535  32 Ice Articuno 580  33 Ice Regice 580  34 Ice Vanilluxe 535  35 Normal Arceus 720  36 Normal Slaking 670  37 Normal Regigigas 670  38 Poison Crobat 535  39 Poison Nidoqueen 505  40 Poison Nidoking 505  41 Psychic Mewtwo 680  42 Psychic Lugia 680  43 Rock Tyranitar 600  44 Rock Diancie 600  45 Steel Dialga 680  46 Steel Metagross 600  47 Steel Jirachi 600  48 Water Palkia 680  49 Water Kyogre 670  50 Water Manaphy 600 |

## Question 2

Imagine a new Pokemon game where you are only allowed to collect ONE type of Pokemon. Similar to other Pokemon games, your goal is to have the strongest battlers and defenders for battles and raids. Which type will you pick? Why?

## Question 3

If you want to predict whether the Pokemon is a legendary Pokemon (a.k.a. predict the field isLegendary using other fields), which models would you use? List your top 3 models with pros and cons for each one.

## Question 4

Pick one model and implement it in a language you are most comfortable with (preferably Python or R). Please do not use the ‘Catch\_Rate’ field (if you are Pokemon fan you know why :). What is your in-sample classification accuracy and what fields did you end up using?

Your answer should include: 1) The code of implementing the model (incl. feature processing, model fitting and cross validating); 2) The formula/description of your final model along with the accuracy number. 3) In addition to the code and the model specification, if you choose to submit a presentation/ dashboard as part of your writeup, you can present your results in any way you like.