A5 Documentation

Author: Shirley Wang

Topic Background and Planning

Since there is a lot of freedom this time in choosing what topic we wanted to explore, I decided to go back to my favourite hobbies and interests. Trading card games don't have any good datasets that aren't just card databases, so I decided on going with something related to Pokemon. Pokemon has datasets detailing all the pokemon, and there was also a poll that happened a few years ago where people voted on their favourite Pokemon, so I could do something related to what sort of Pokemon people like. The topic detailed in my original plan was to investigate if lacking Pokemon in a new Pokemon game actually affected the ratings and sales of a game, from the controversy online a few months ago about the lack of Pokemon in the new Pokemon game, but from feedback from that tutorial the scale of that analysis is a bit too big for one poster, plus that sort of analysis is hard to do from just data. So I decided on a different controversy that I could try and tackle with the data.

Main Points:

- Generation 1 Pokemon are the most popular
- The popularity of Pokemon did decline from Generation 5 and beyond
- Older pokemon are liked more because they are older, not because of other factors.

There is a very loud group of people online who believe that the first generation of Pokemon, known as Gen 1 Pokemon, are the best, and one of the most common arguments for this is that new Pokemon have bad designs as compared to the old ones. It is true that a lot of older Pokemon are a lot more popular than new ones, Pikachu being basically famous worldwide as an example, but this group online is very vocal about how newer games are all much worse and generation 1 Pokemon are much better, and it's annoying to be around.

My idea is that if Pokemon from newer generations are similar to Pokemon from Gen 1 but have significantly less popularity votes, then that would imply a bias for Gen 1 induced from nostalgia from the voters, rather than the Pokemon in Gen 1 actually being better than later generations. So I would like to focus my poster on showing that generation appears to be a big factor in what determines a pokemon's popularity, to show that there is a bias towards older pokemon rather than newer ones for popularity.

Data

The rounakbanik kaggle Pokemon dataset contains basic information about all Pokemon from generations 1 to 7, except 8 legendary pokemon that were released in Ultra Sun and Ultra Moon and later, so I have manually gone into the file and added in their data. Meltan and Melmetal are weirdly split between Generation 7 and 8, so I'm putting them in Generation 7 cause I don't have any vote data for anyone else in Generation 8.

The main measure of popularity I'll be using comes from a reddit survey from 2019 where a voter would answer what their favourite Pokemon was, from generations 1 to 7. Over 52000 votes were counted, and so each Pokemon from generations 1 to 7 has a number of votes corresponding to how many people think of it as their favourite Pokemon. I will be using both the number of votes a Pokemon and their respective rank from the votes as measures of how well liked a Pokemon is.

The thing about using the reddit survey votes as a measure of popularity is that the dataset is definitely biased towards older fans of Pokemon, specifically fans of Pokemon who also spend a lot of time on reddit,

but conviniently a lot of the irritating people that originated this problem I'm investigating are also on reddit, so this dataset should specifically represent the group of people I am interested in investigating.

Design

A rough outline of answers to the questions in the handout.

- Audience: people who like Pokemon.
- Data: covered above.
- Insights about Data: I would like to show that there is a bias to like older Pokemon over newer pokemon. I'd also like to discover what kind of pokemon are popular.
- How does your statistical analysis support these insights? Hopefully I'll find some nice presentable numbers. I will be using Poisson regression for finding useful factors, although I'm worried about how presentable the results will be on a poster.
- Why is this important? Pokemon is a pretty popular game, so it's important at least hobby-wise to a lot of people.
- What is the story that you are going to tell? I would like it to go something like this
    1. Show popularity of pokemon by generation
    2. Show specific similar pokemon in different generations with popularity votes (I'll try and see if I can visualize this in interesting ways)
    3. See if I can find any other factors that lead to high popularity (this may not make it into the final poster)
    4. Conclude that a lot of what has a pokemon be popular has to do with what generation it came out in, rather than design choices.
- What effect are you hoping to have? I think it would be nice if a Gen One-ner looked at this poster and went "hmm maybe I think Gen 1 is better is because I'm a jaded adult now and don't like new things, not because new things are actually worse. I should stop being so irritating online."

Main Visualizations

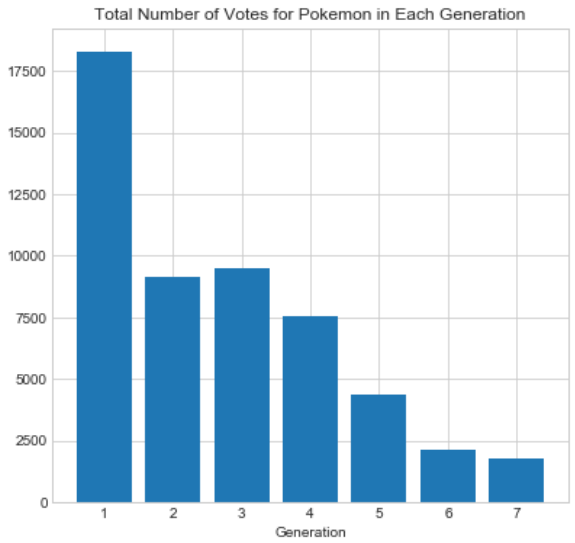
The three perspectives I've decided on for the main visualizations of the poster is to go from large to small. 1. How Gen 1 pokemon have the most votes in total / on average 2. How within each Pokemon type, the majority of them have the highest average votes for Gen 1 3. How for a specific set of very similar Pokemon that were released in different generations, the Gen 1 pokemon has the highest number of votes.

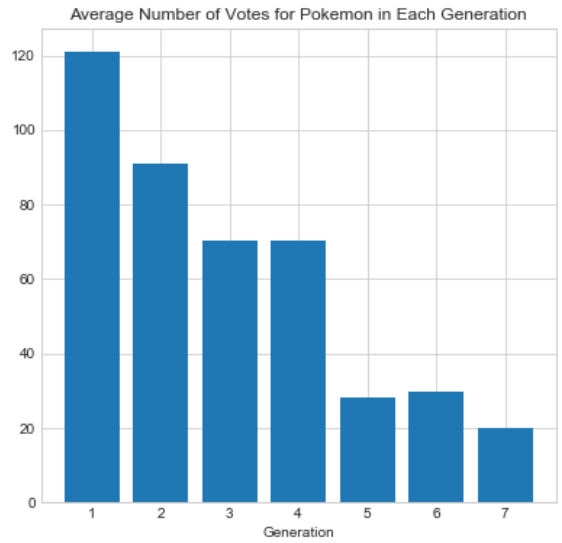
Part 1: Aggregated Generation Votes

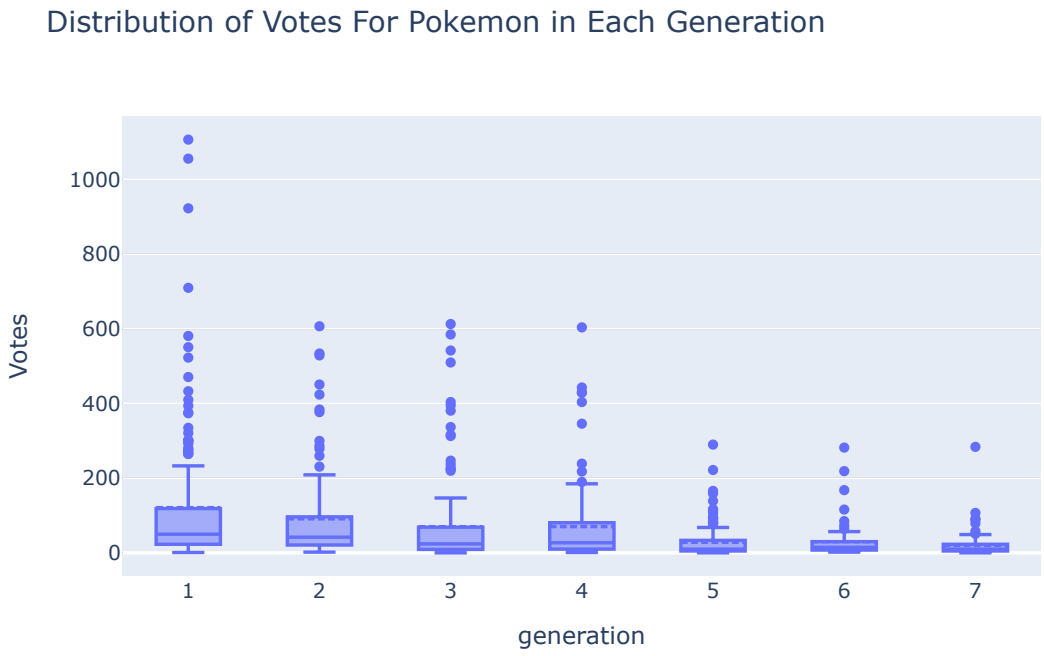
Perspective #1: Generation 1 Pokemon on average are more popular than other pokemon.

Preliminary Data Analysis

Some visualizations showing how the generations compare.







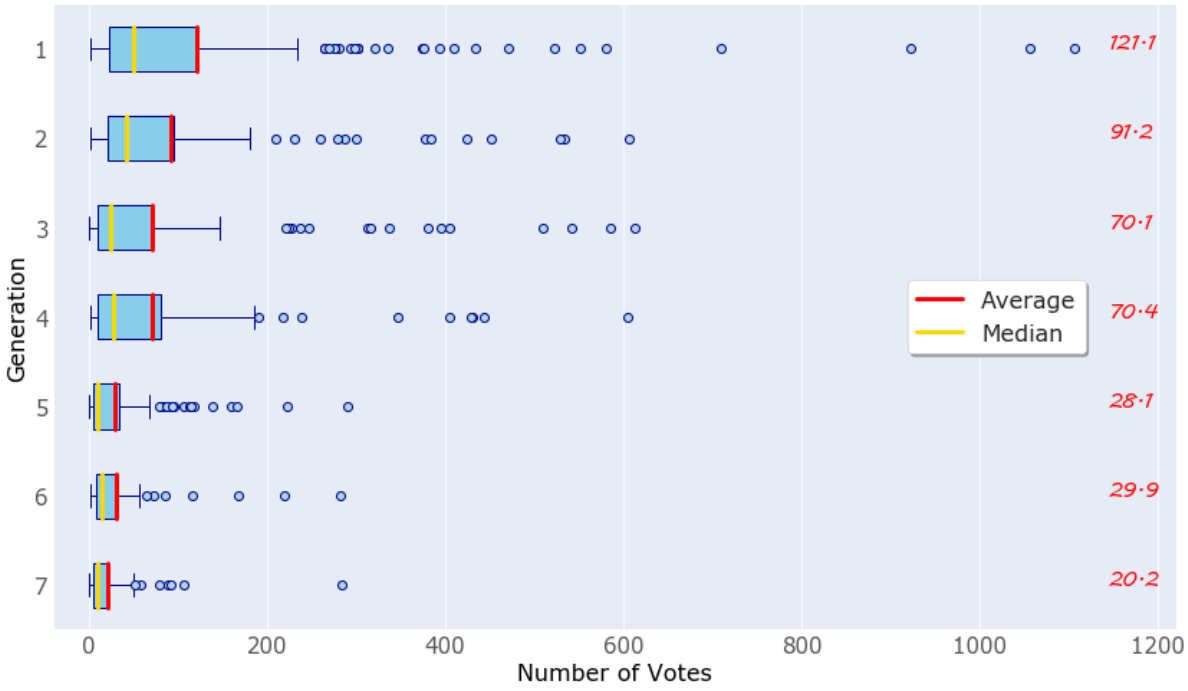
From these we can see that there are 4 extremely popular pokemon in Generation 1: Charizard, Gengar, Arcanine, and Bulbasaur (in order). After those four, the distribution of votes for generations 1 to 4 are relatively similar, although generation 1 does still have a slightly higher boxplot that the others. Generations 5 to 7 pokemon are pretty unpopular in comparison. There is actually a pretty big opinion online that Gen 5 and after is when Pokemon "started going downhill", so this does reflect that.

Main Visualization Idea

Average number of votes per pokemon for each generation would be better than total votes, because the number of pokemon in each generation varies quite a lot. Generation 1 has 151, while Generation 6 has 72 pokemon, so taking the average adjusts for this. Showing the distribution of votes would also be helpful, since it'll show how Generation 1 also has the Pokemon that have the highest votes. The nice thing about plotly is above if you hover your mouse over any of the outliers, you can see which pokemon it is that has such a higher number of votes, which I can't do on a poster. I considered labelling a few outliers, but it would make the visualization a bit too crowded so I didn't.

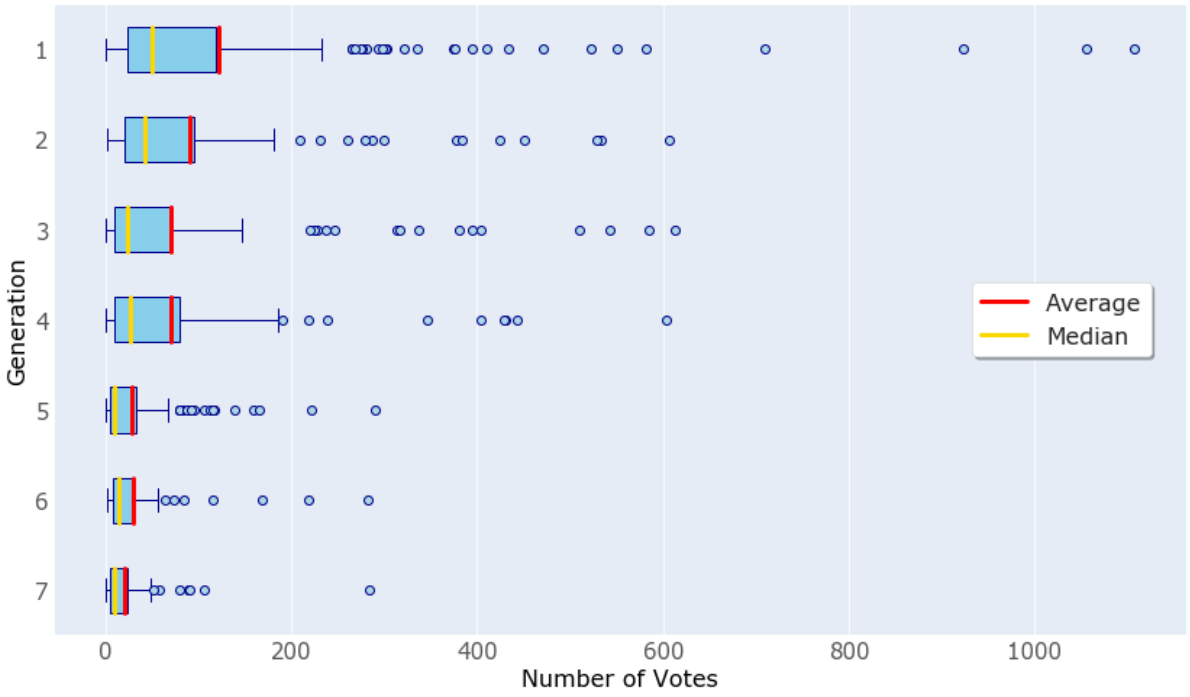
Version 1

Plotly is a bit too limiting. Matplotlib does allow for more freedom in the customizations, so I can go ahead and make the specifics of what I want. I also decided having the titles made in the poster is better than using the matplotlib ones.



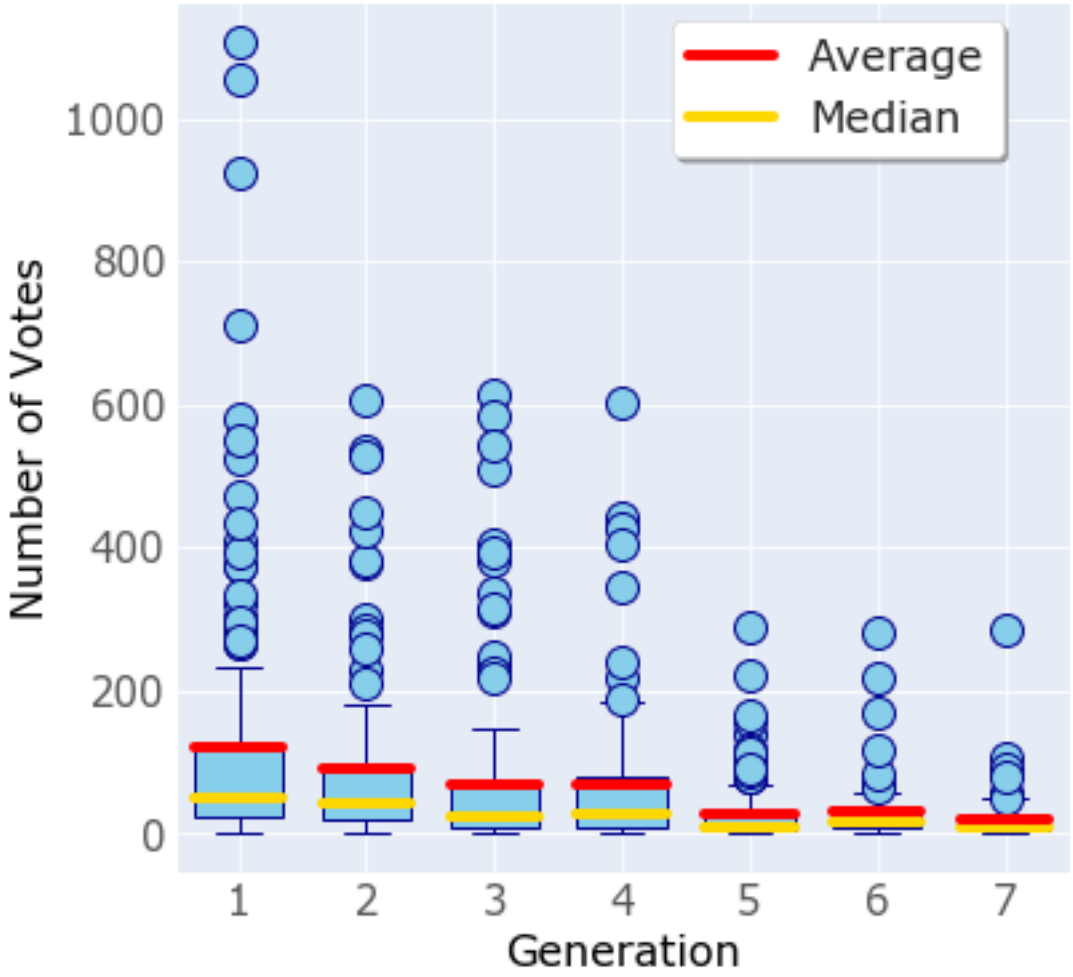
Version 2

I thought adding the averages on the side would help emphasize that the averages seem to be decreasing as generation increases, but it looks kinda awkward. Version 2 doesn't have the averages.



Version 3

After I made my first two drafts of the poster and I wanted to adjust it a bit. Vertical works better for its position on the poster. This is the final version in the poster.



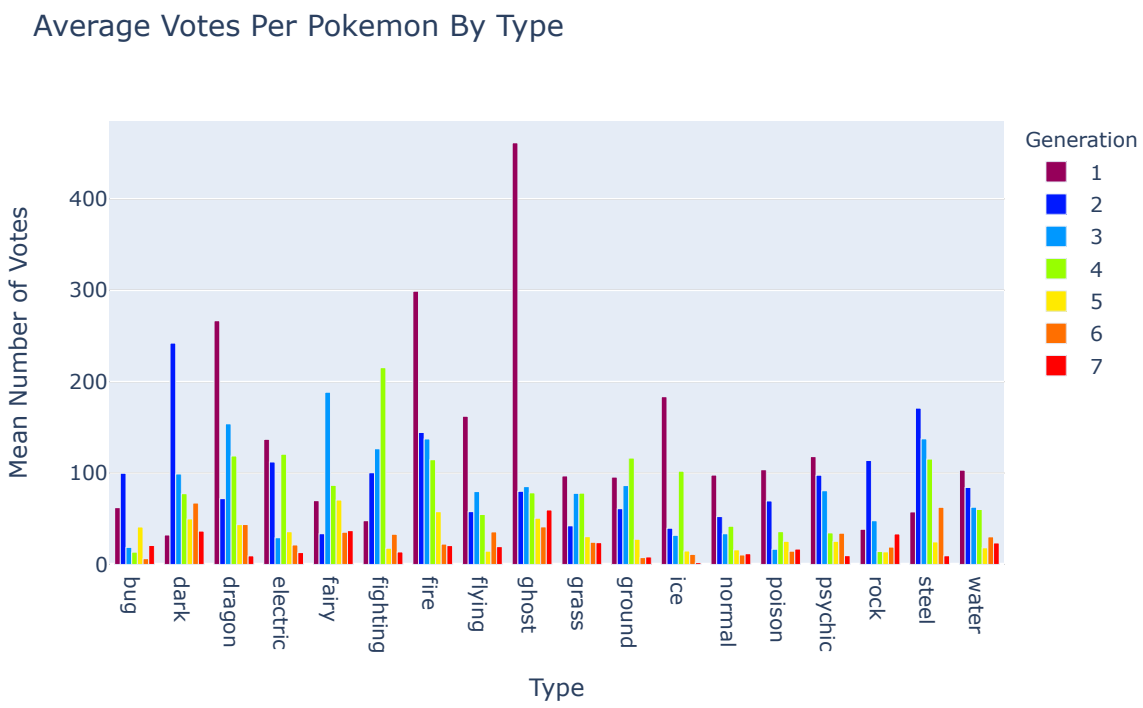
Part 2: Breakdown of Votes within A Category

Perspective 2: Within each Pokemon type, most of the time Generation 1 pokemon have on average more votes. As well, Generation 5 and beyond have the highest average for none of the types.

I decided to further view the votes distribution by Pokemon Type, since a lot of people tend to have a favourite type, and some pokemon types are more popular than others.

Version 1

Viewing it in plotly to get a better idea of how it looks.



For a lot of these types the average number of votes for each type is highest in generation 1. 11/18 types have generation 1 having the highest mean. Exceptions are Bug, Dark, Fairy, Fighting, Ground, Rock, and Steel.

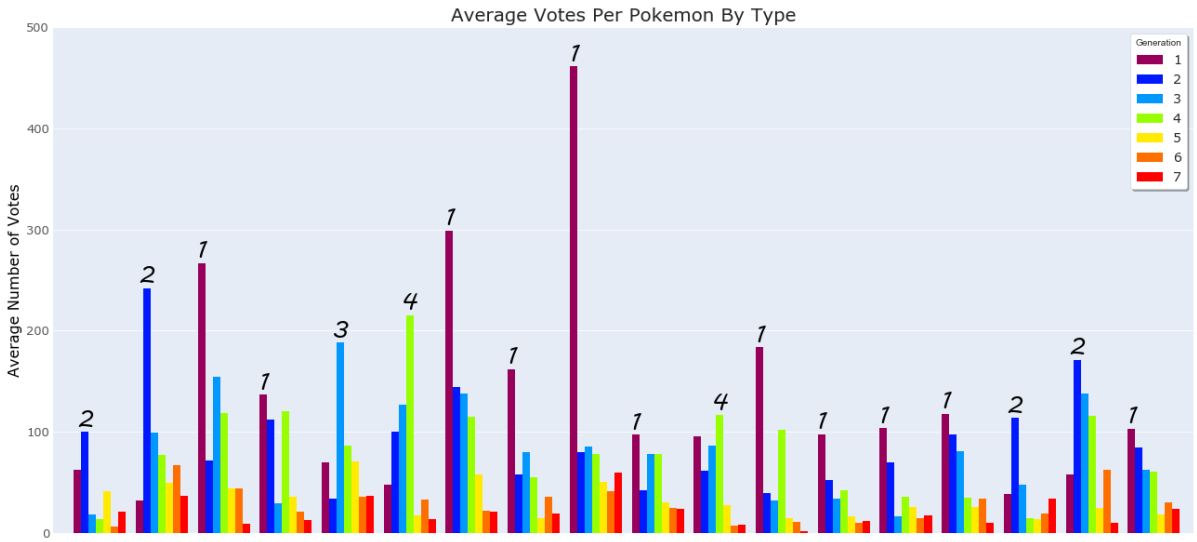
Dark, Steel, Fairy: - Dark and Steel type aren't introduced until generation 2, so only a few pokemon in generation 1 were later reclassified as these types. That also explains why for these types the highest mean is in generation 2. - Fairy type is technically introduced in generation 6, but it involves a remake of the generation 3 main game that made and featured many of them as fairy types, so it makes sense that generation 3 has the highest mean for fairy types.

Actual Outliers: - Bug (Gen 2) - Fighting (Gen 4) (Lucario (lots of marketing) was in this gen so that might be why) - Ground (Gen 4) (Garchomp (very popular) was in this gen so that might be why) - Rock (Gen 2)

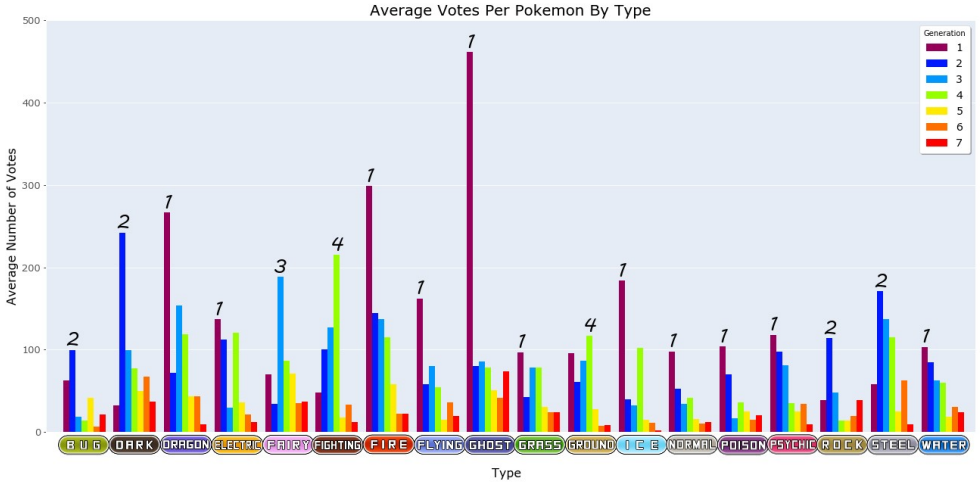
Version 2

With numbers above each type representing which generation has the highest mean. It doesn't look terrible. My plan is to also replace the x axis with the official pokemon type stamps.

I had to remake the thing in Matplotlib just for formatting consistencies. After making it in matplotlib I went into a photo editor and added in the official logos for types.

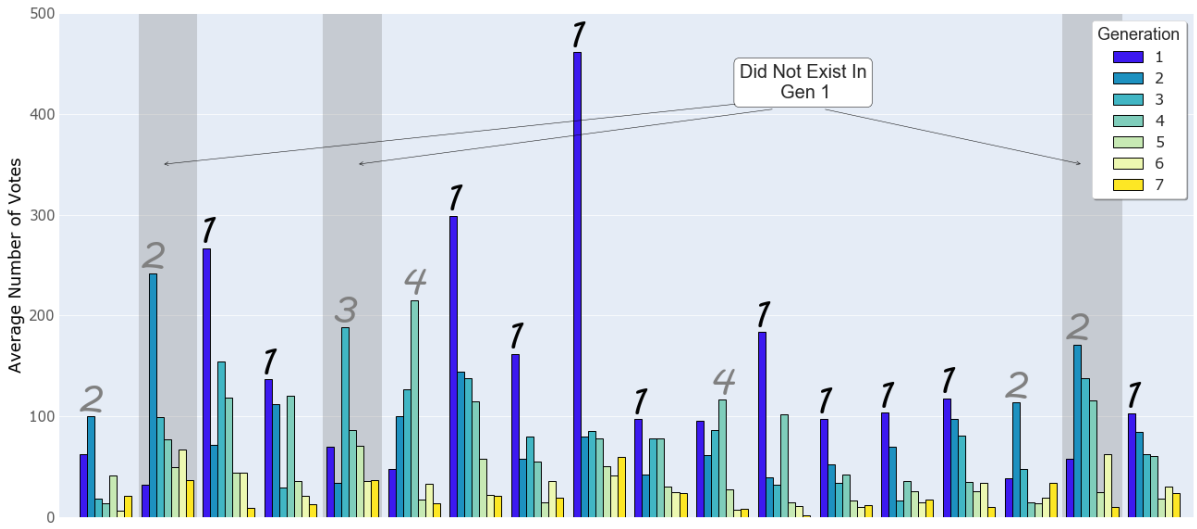


And then I realized that making both the bars and the types colorful may be a bit too much.

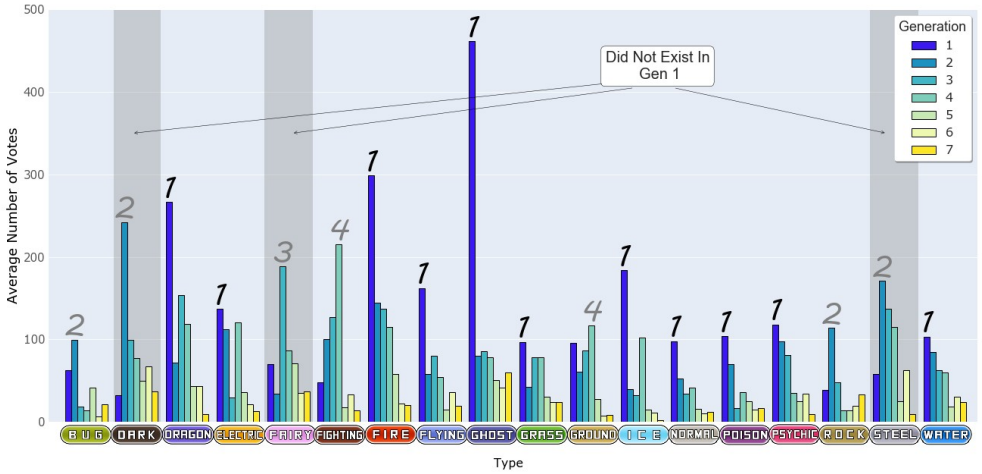


Version 3

I decided on using a Blue to Yellow color scheme, to match the Pokemon logo. By recommendation from Michael, I also added in an indicator for the three types that weren't introduced until after generation 1.



Poster's final version:



Part 3a: Finding Significant Factors that affect Popularity

I used Poisson regression with Votes as the response and a few predictors to try and find which factors seemed to influence the number of votes a pokemon would have. I'm using Poisson instead because number of votes is an integer count, and right-skewed like the Poisson distribution. I also use $\alpha = 0.001$ instead of the usual 0.05 because there are so many predictors, so I want to be really confident about which types have effects. Similar to A1, I'm trying to make a point that generation is the most important predictor, like a confounding variable. As well, I also want to find a few other good predicting factors, so that I can compare similar pokemon across generations and show that there is a bias towards older ones later.

Each Pokemon can have either 1 or 2 types. Since the distinction between type1 and type2 is nonexistent (if one Pokemon has type1 = "Fire" and another has type2 = "Fire", they're both Fire types), I'm recoding the two columns to a set of dummies for each type, rather than a set of dummies for types in each column. I also have a single type flag, to check if single or double typed pokemon are more popular

or not. People tend to have favourite types, and some types are more popular than others, so this is to check that.

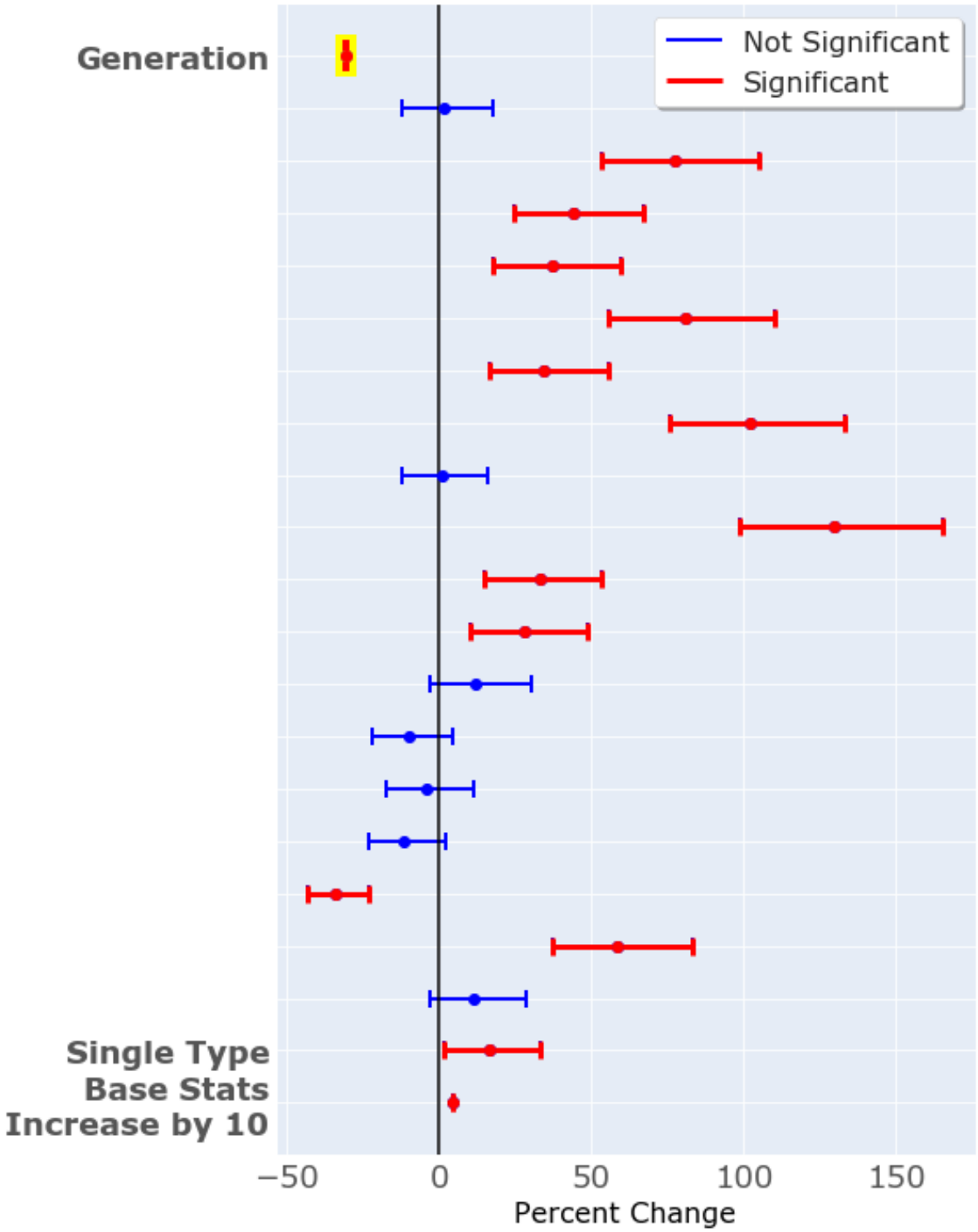
Base total represents the sum of their six base stats (HP, Attack, Special Attack, Defense, Special Defense, Speed), and is a simple measure of how strong a Pokemon is. The lowest base stat total is 180 for Sunkern (a small funny grass type pokemon), and the highest is 780 (in this dataset at least), and the increase by 1 in base stat total honestly isn't that noticable to anyone. So I've normalized the base stats total by subtracting 180 and then dividing it by 10 so the percent change is easier to see and understand.

Model 1 is the model without Generation added as a predictor, and Model 2 is the model where I added Generation as a predictor.

	1: coef	2: coef	1: std err	2: std err	1: Sig	2: Sig	Consistent
Intercept	3.1324	3.7269	0.083	0.083	*	*	
Is_Bug	-0.3588	0.0157	0.045	0.045	*		
Is_Dark	0.1356	0.5735	0.043	0.044		*	
Is_Dragon	-0.2540	0.3653	0.044	0.044	*	*	
Is_Electric	0.0559	0.3143	0.045	0.046		*	
Is_Fairy	-0.1063	0.5925	0.045	0.045		*	
Is_Fighting	-0.1859	0.2973	0.044	0.044	*	*	
Is_Fire	0.4784	0.7050	0.043	0.043	*	*	*
Is_Flying	-0.0500	0.0081	0.043	0.043			
Is_Ghost	0.1198	0.8318	0.044	0.044		*	
Is_Grass	-0.1595	0.2847	0.044	0.044	*	*	
Is_Ground	0.0287	0.2465	0.044	0.045		*	
Is_Ice	-0.0149	0.1148	0.045	0.045			
Is_Normal	-0.3757	-0.1025	0.044	0.044	*		
Is_Poison	0.1404	-0.0438	0.045	0.045			
Is_Psychic	-0.2873	-0.1210	0.043	0.044	*		
Is_Rock	-0.6341	-0.4167	0.046	0.046	*	*	*
Is_Steel	-0.0436	0.4609	0.044	0.044		*	
Is_Water	-0.0583	0.1099	0.042	0.043			
Single_Type	-0.0248	0.1525	0.041	0.041		*	
base_total_normalized	0.0420	0.0438	0.000	0.000	*	*	*
generation	NaN	-0.3685	NaN	0.003		*	

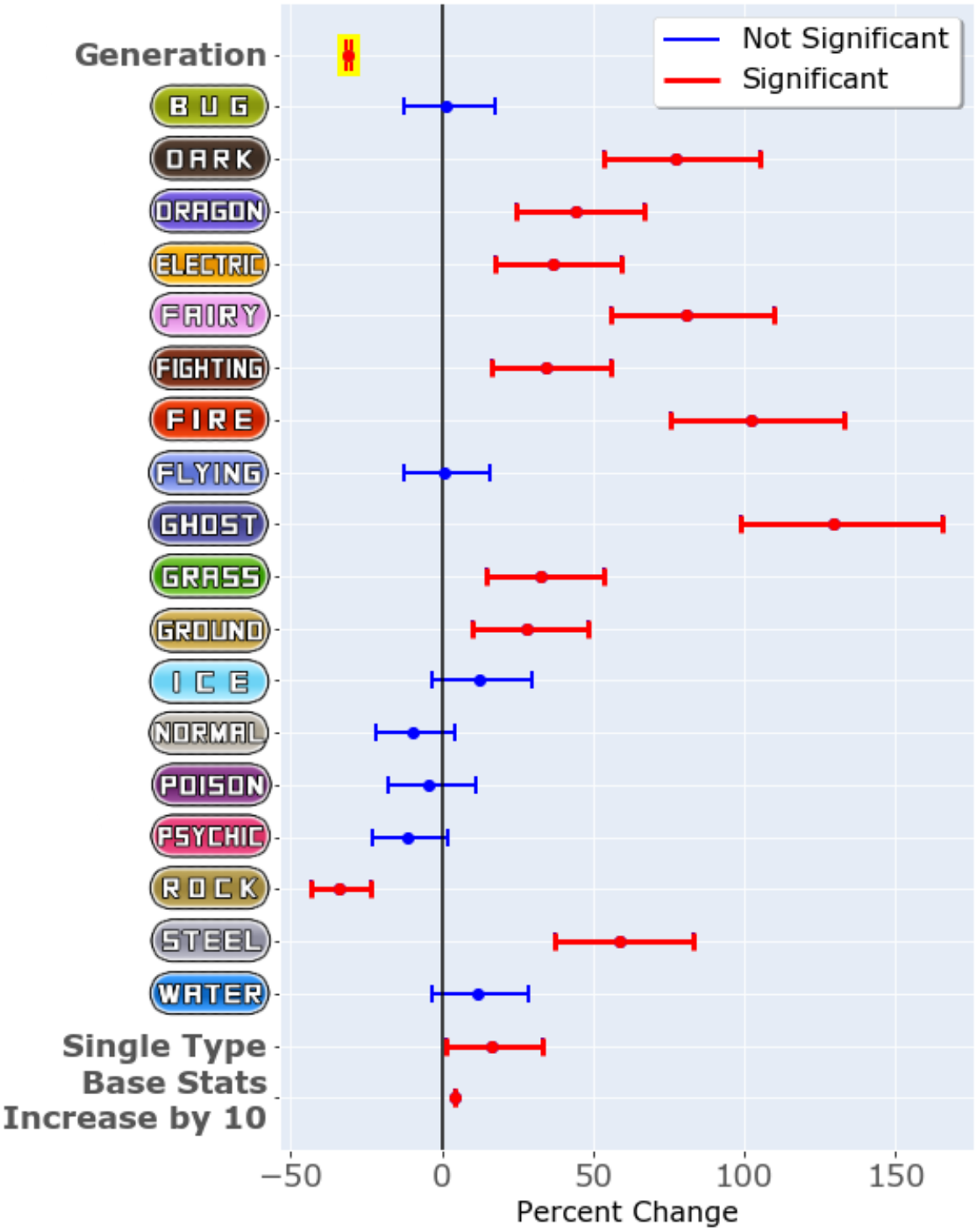
A "consistent" effect is one where the confidence intervals in Model 1 and 2 overlap, so adding Generation hasn't changed it too much. We can see that barely anything overlap, and there are barely any consistent effects. Basically, the predictions for the effects of most types significantly change when Generation is added into the model, so it's very likely Generation is a confounding variable.

From now on out I'll be using Model 2 cause it has Generation included in it.



I did try to highlight how Generation's effect was really negative by putting it at the top and putting yellow around it, although it doesn't look the best in my opinion.

Final version:



Part 3b: Example of Gen 1 Bias

Perspective #3: A specific group of very similar pokemon across generations has the highest popularity at the generation 1 pokemon.

There is a group of Pokemon known as pseudo-legendary pokemon. They all have a base stat total of 600, are all Dragon type except for two (but the two exceptions have types Dark and Steel, two other types

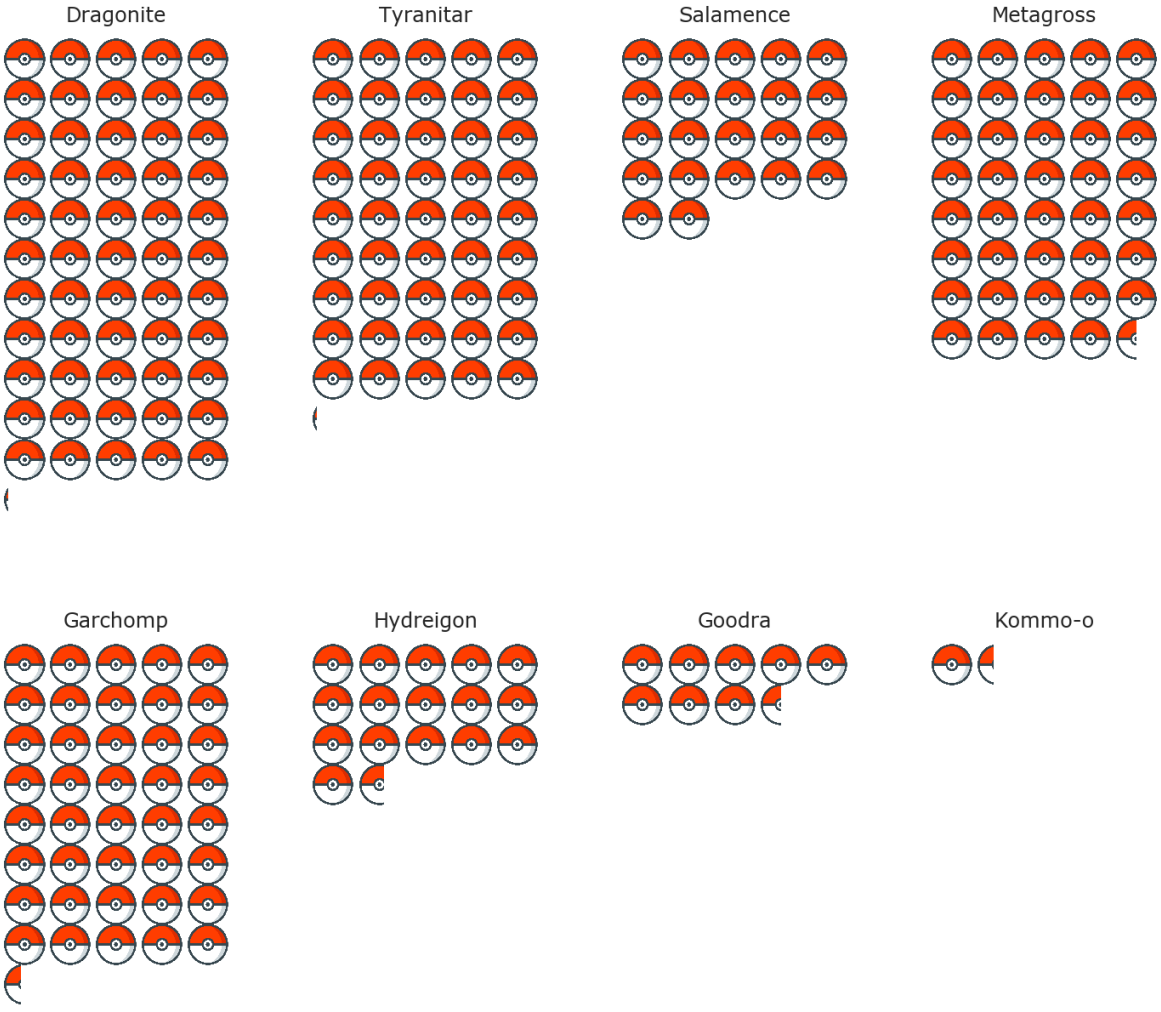
that are expected to have significant positive effects). They also all have relatively the same amount of marketing around them, since all of them are featured in their debut games as Pokemon belonging to one of the strongest trainers in the game. So this seems like a good group to study since a lot of other factors around the Pokemon are controlled.

As expected, Dragonite has the highest number of votes despite all of the pokemon having relatively the same factors, which backs up my point that there is nostalgia bias in liking a pokemon.

I was originally gonna put pictures of the pokemon on top of the bars, but then I realized that Part 2 was also a bar plot, so I wanted to do something different for this one. I imaged a pictogram using pokeballs to represent votes cause it was thematically appropriate and funny. So I made my own.

Version 2

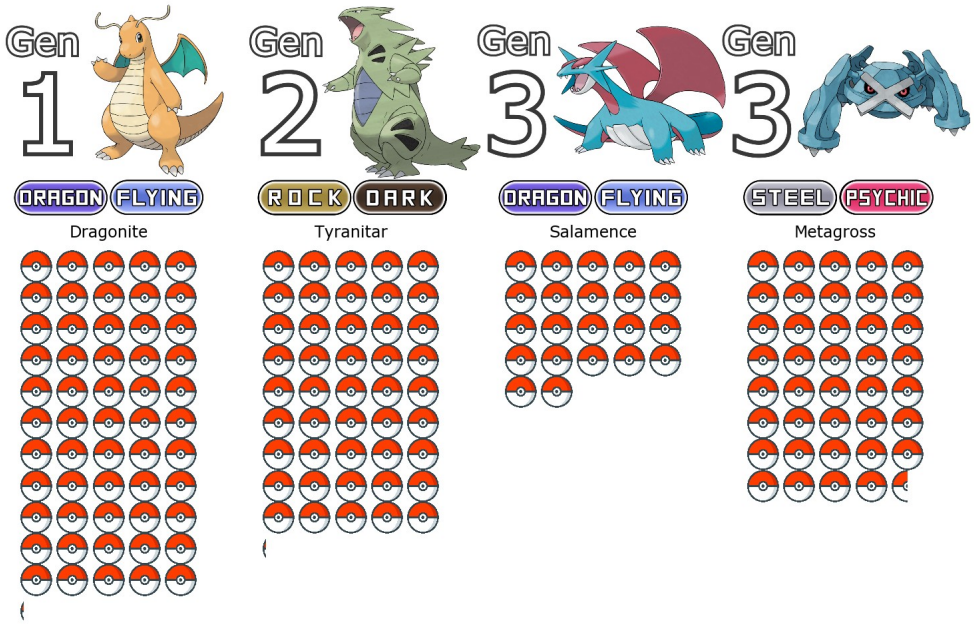
Using 1 pokeball = 10 votes.



This is everything I can do in matplotlib. The rest I did in an image editor.

The legend:

The final version:


Gen 1
DRAGON FLYING
Dragonite
Gen 2
ROCK DARK
Tyranitar
Gen 3
DRAGON FLYING
Salamence
Gen 3
STEEL PSYCHIC
Metagross
Gen 4
DRAGON GROUND
Garchomp
Gen 5
DRAGON DARK
Hydreigon
Gen 6
DRAGON
Goodra
Gen 7
DRAGON FIGHTING
Kommo-o
= 10 Votes

Poster Design

Version 1

There's an empty space at the bottom and the color of the figures don't really fit together the best.

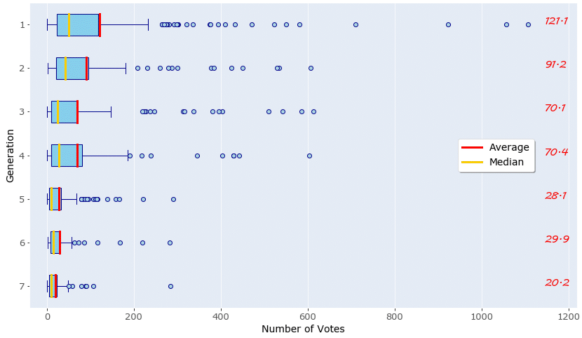
Generation 1 Nostalgia and Bias

It is a common opinion among some very vocal older Pokemon fans that Generation 1 (the original games Pokemon Red and Pokemon Blue) had the best story, the best themes, and the best pokemon. While this group of people are typically not taken too seriously, their mindset may be more widespread than initially thought.

Using the results from the Pokemon popularity survey held in 2019, where each person would cast a vote for their favourite pokemon, it becomes clear that older generation pokemon seem to be more popular than newer ones, and that lots of people are prone to letting nostalgia influence their preferences.

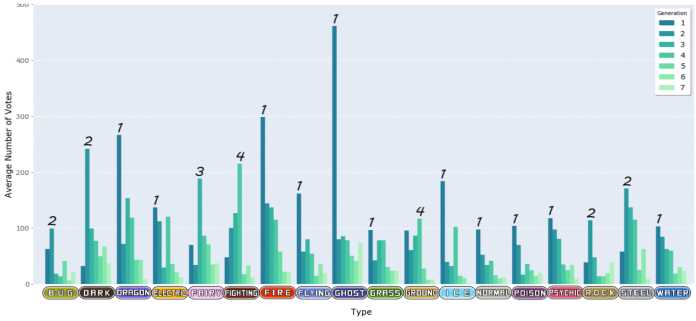
Votes Per Pokemon By Generation

Generation 1 has the highest average votes per pokemon, the highest median votes per pokemon, and the four pokemon with the highest number of votes.



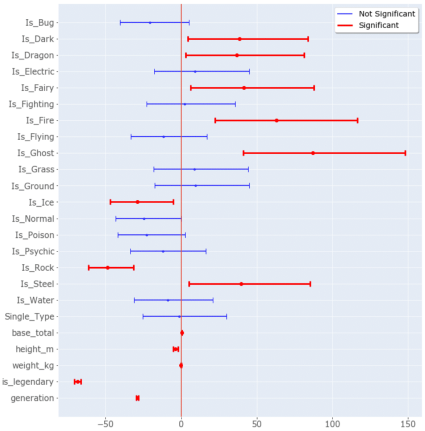
Average Votes Per Pokemon By Type

Generation 1 has the highest average votes per pokemon for 11/18 types. 3 types that it doesn't have the highest average in weren't introduced until later games. The types Generation 1 doesn't have the highest average in has either Gen 2, 3, or 4 as their highest, also older generations.



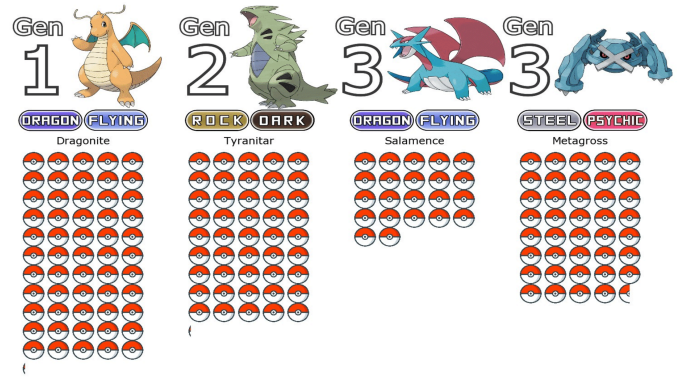
Expected Percent Change in Number of Votes

Running a Poisson Regression to model the votes and popularity a pokemon will have, we get that Generation has a significantly negative effect on votes for every 1 unit increase.



Similar Pokemon from Different Generations

Pseudo-legendary pokemon are 8 pokemon from the first 7 generations that all have a base stat total of 600, are almost all Dragon type (the two exceptions are Dark and Steel type, two more types that expect a similar positive influence on number of votes), and all have a similar level of marketing (being featured in their respective games as belonging to one of the strongest pokemon trainers). Despite this, Generation 1's Dragonite has the most votes.



Additional Factors Affecting Popularity

The model predicts a positive relationship between base stat total of a Pokemon and number of votes it would get.

Types with Positive Effects:

- Dark
- Dragon
- Fairy
- Fire
- Ghost
- Steel

Types with Negative Effects:

- Ice
- Rock

Keeping An Open Mind

The people that grew up with the older Pokemon games are now all old enough to complain on the internet about the new ones. The childlike wonder that came from playing that first game is long gone, but that doesn't mean that new games are bad because of it. It is understandable that many people have a preference for the pokemon they grew up with, but keeping an open mind in the future will allow them to continue to enjoy pokemon without arguing with others online.

Version 2

I did fill up the empty space by just kinda stretching things and repeating what I said for the regression portion. I found this version a bit too wordy. Plus after the presentations on Tuesday I thought mine looked kinda ugly compared to some of the others.

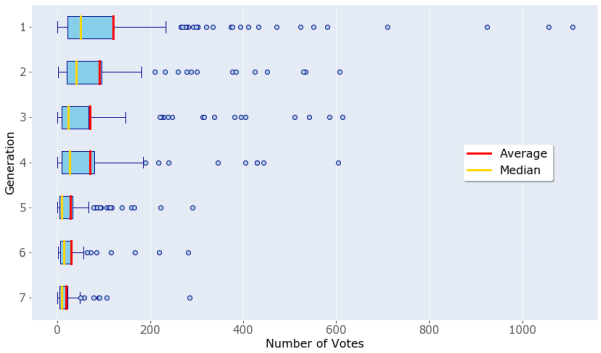
Generation 1 Nostalgia and Bias

It is a common opinion among some very vocal older Pokemon fans that Generation 1 (the original games Pokemon Red and Pokemon Blue) had the best story, the best themes, and the best pokemon. While this group of people are typically not taken too seriously, their mindset may be more widespread than initially thought.

Using the results from the Pokemon popularity survey held in 2019, where each person would cast a vote for their favourite pokemon, it becomes clear that older generation pokemon seem to be more popular than newer ones, and that lots of people are prone to letting nostalgia influence their preferences.

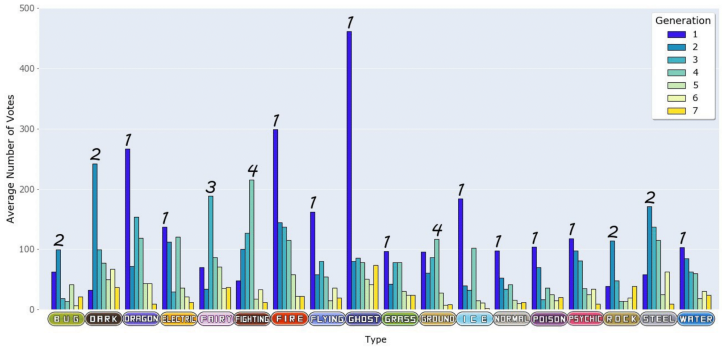
Votes Per Pokemon By Generation

Generation 1 has the highest average votes per pokemon, the highest median votes per pokemon, and the four pokemon with the highest number of votes.



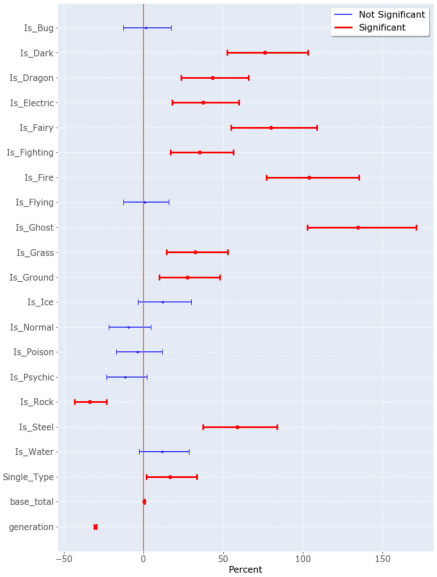
Average Votes Per Pokemon By Type

Generation 1 has the highest average votes per pokemon for 11/18 types. 3 types that it doesn't have the highest average in weren't introduced until later games. The types Generation 1 doesn't have the highest average in has either Gen 2, 3, or 4 as their highest, also older generations.



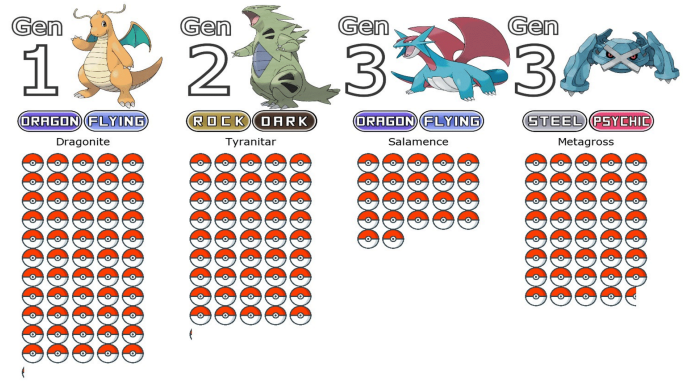
Expected Percent Change in Number of Votes

Running a Poisson Regression to model the votes and popularity a pokemon will have, we get that Generation has a significantly negative effect on votes for every 1 unit increase.



Similar Pokemon from Different Generations

Pseudo-legendary pokemon are 8 pokemon from the first 7 generations that all have a base stat total of 600, are almost all Dragon type (the two exceptions are Dark and Steel type, two more types that expect a similar positive influence on number of votes), and all have a similar level of marketing (being featured in their re-



Additional Factors Affecting Popularity

The model predicts a positive relationship between base stat total of a Pokemon and number of votes it would get, and a negative relationship between generation of the pokemon and number of votes.

Types with Positive Effects:

- Dark
- Dragon
- Electric
- Fairy
- Fighting
- Fire
- Ghost
- Grass
- Ground
- Steel

Types with Negative Effects:

- Rock

Keeping An Open Mind

The people that grew up with the older Pokemon games are now all old enough to complain on the internet about the new ones. The childlike wonder that came from playing that first game is long gone, but that doesn't mean that new games are bad because of it. It is understandable that many people have a preference for the pokemon they grew up with, but keeping an open mind in the future will allow them to continue to enjoy pokemon without arguing with others online.

Version 3

I tried to get rid of a lot of the words and also increase the text size for better readability. I also readjusted the placement of the graphs, gave more room for the title. Not sure the conclusion portion looks the best, but this is the best poster I can make so far.

Older Fans, Older Generation Bias

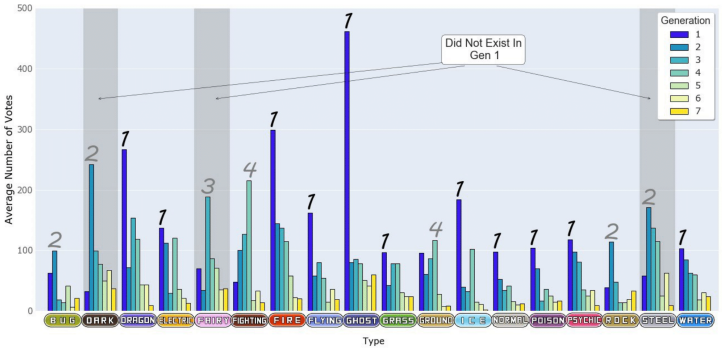
There is a vocal group of fans online who believe **Generation 1** had the **best Pokemon**.

In reality, **most older fans** are biased towards Pokemon from older generations, especially **Generations 1 to 4**.

In 2019, **52000+** people on Reddit voted for their favourite Pokemon.

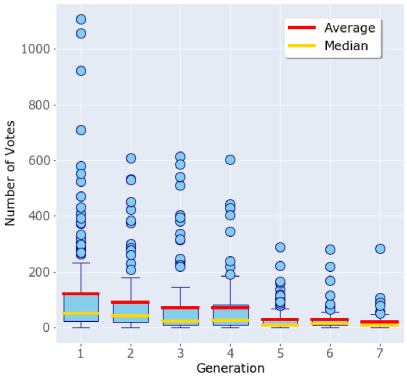
Average Votes Per Pokemon By Type

Generation 1 has the **highest average votes** per Pokemon for **11/18 types**. All types have their highest average in **Generations 1 to 4**.



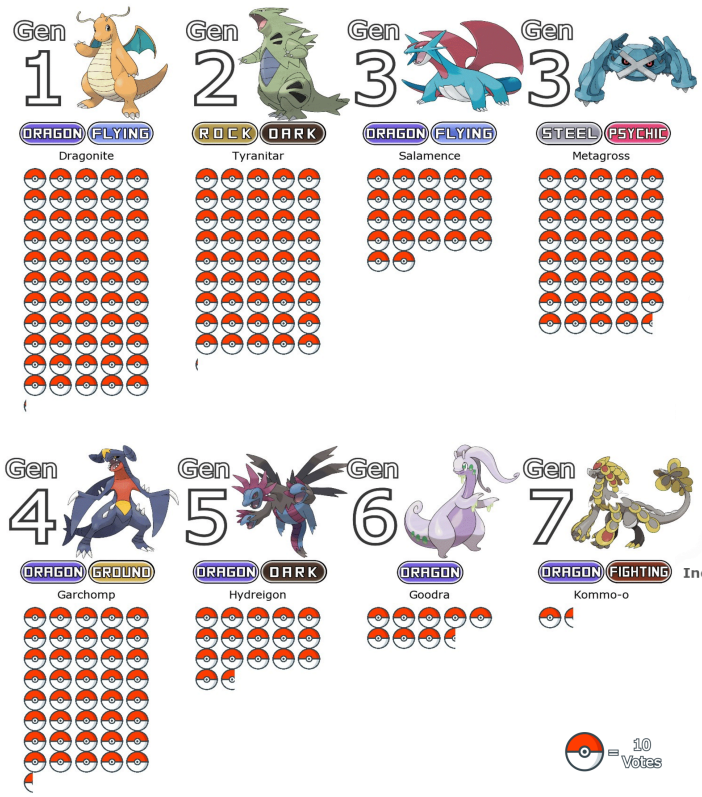
Votes By Generation

Generation 1 has the **highest average votes** per Pokemon, **highest median votes**, and four Pokemon with the **highest number of votes**.



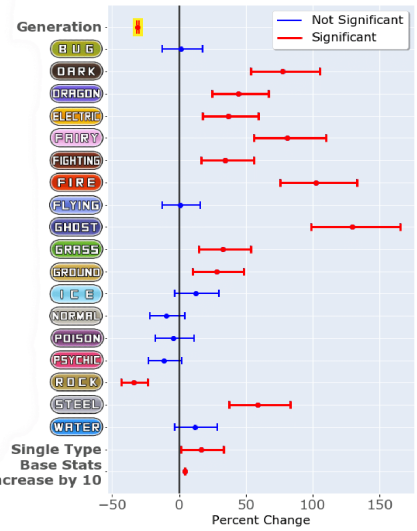
Similar Pokemon from Different Generations

Pseudo-Legendary Pokemon all have a **base stat total of 600**, have **similar types**, and have a **similar level of marketing**. **Generation 1** still has the most votes.



Predicted Percent Change in Votes

A statistical model was fit to predict the votes of a Pokemon. Generation has a **significantly negative effect on votes** for every increase.



Newer Pokemon Are *Not* Worse

Older fans are just nostalgic for their childhoods and biased towards older generation Pokemon.

Additional Analysis

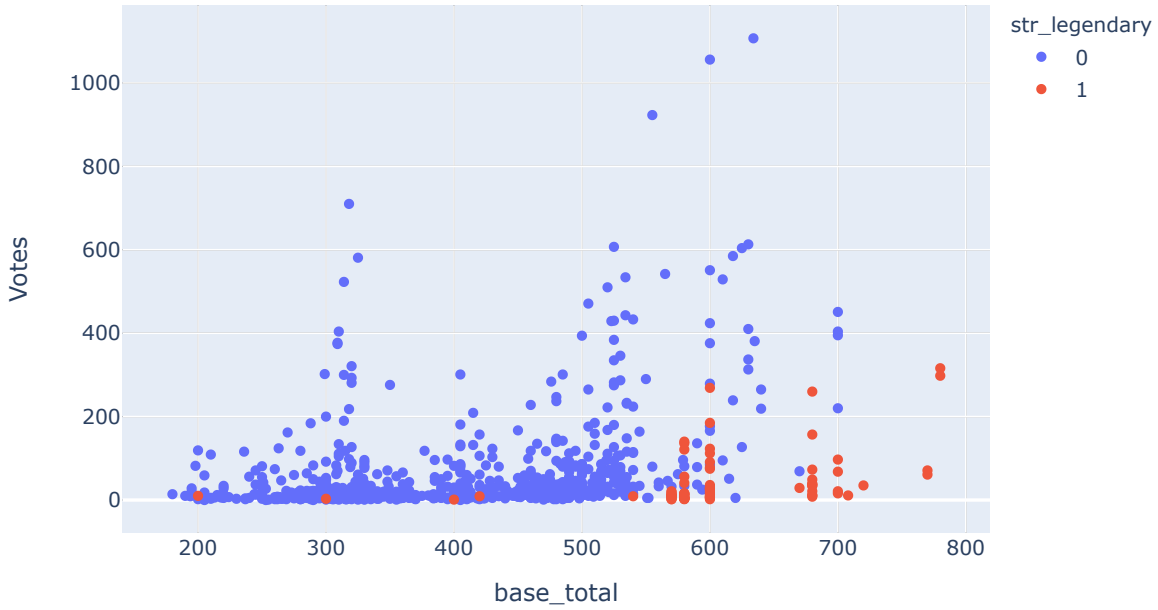
Some more findings that didn't make it into the final poster.

Poisson Regression with More Factors

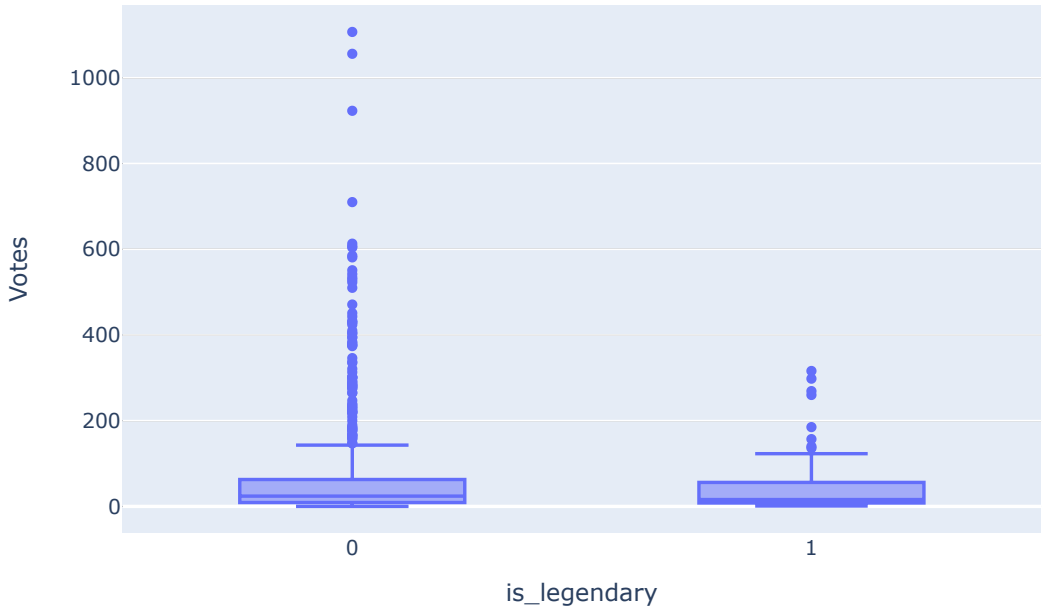
	coef	std err	z	P> |z|	[0.0005	0.9995]
Intercept	3.7154	0.170	21.841	0.000	3.156	4.275
Is_Bug	-0.2379	0.086	-2.755	0.006	-0.522	0.046
Is_Dark	0.3225	0.085	3.774	0.000	0.041	0.604
Is_Dragon	0.3080	0.086	3.582	0.000	0.025	0.591
Is_Electric	0.0814	0.086	0.943	0.346	-0.203	0.366
Is_Fairy	0.3416	0.086	3.951	0.000	0.057	0.626
Is_Fighting	0.0192	0.086	0.224	0.823	-0.263	0.301
Is_Fire	0.4779	0.087	5.518	0.000	0.193	0.763
Is_Flying	-0.1261	0.085	-1.480	0.139	-0.406	0.154
Is_Ghost	0.6142	0.086	7.164	0.000	0.332	0.896
Is_Grass	0.0784	0.086	0.908	0.364	-0.206	0.363
Is_Ground	0.0838	0.086	0.975	0.329	-0.199	0.367
Is_Ice	-0.3469	0.088	-3.945	0.000	-0.636	-0.058
Is_Normal	-0.2897	0.086	-3.366	0.001	-0.573	-0.007
Is_Poison	-0.2609	0.086	-3.033	0.002	-0.544	0.022
Is_Psychic	-0.1351	0.086	-1.580	0.114	-0.416	0.146
Is_Rock	-0.6686	0.087	-7.670	0.000	-0.955	-0.382
Is_Steel	0.3292	0.086	3.825	0.000	0.046	0.612
Is_Water	-0.0973	0.085	-1.140	0.254	-0.378	0.184
Single_Type	-0.0182	0.084	-0.216	0.829	-0.296	0.260
base_total_normalized	0.0610	0.001	120.763	0.000	0.059	0.063
generation	-0.3420	0.003	-116.880	0.000	-0.352	-0.332
is_legendary	-1.1574	0.020	-58.287	0.000	-1.223	-1.092
height_m	-0.0384	0.005	-7.218	0.000	-0.056	-0.021
weight_kg	-0.0004	6.03e-05	-7.413	0.000	-0.001	-0.000

Other factors affecting popularity

From the suggestion of an upward trend in votes for base stat totals.

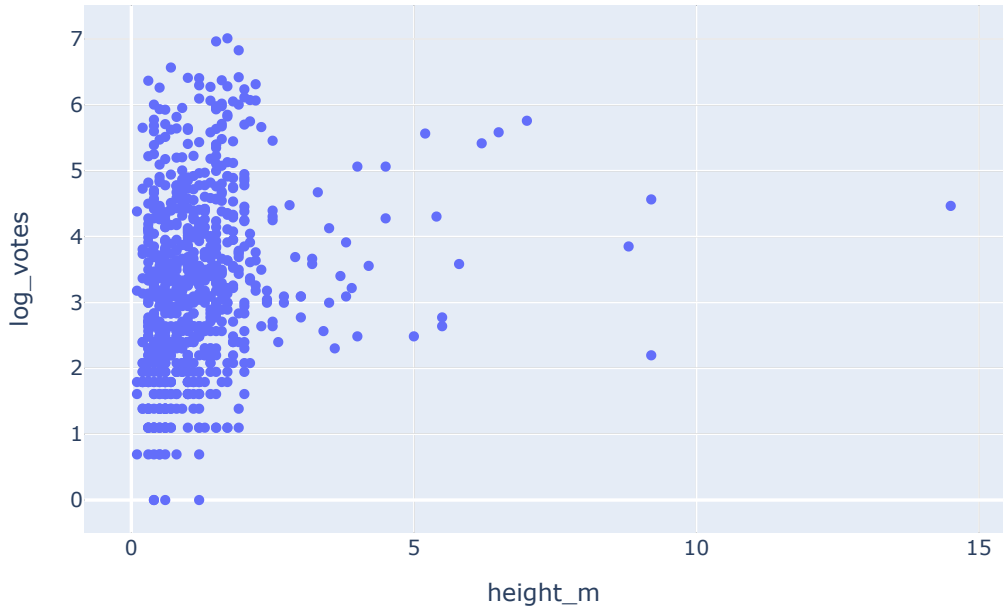


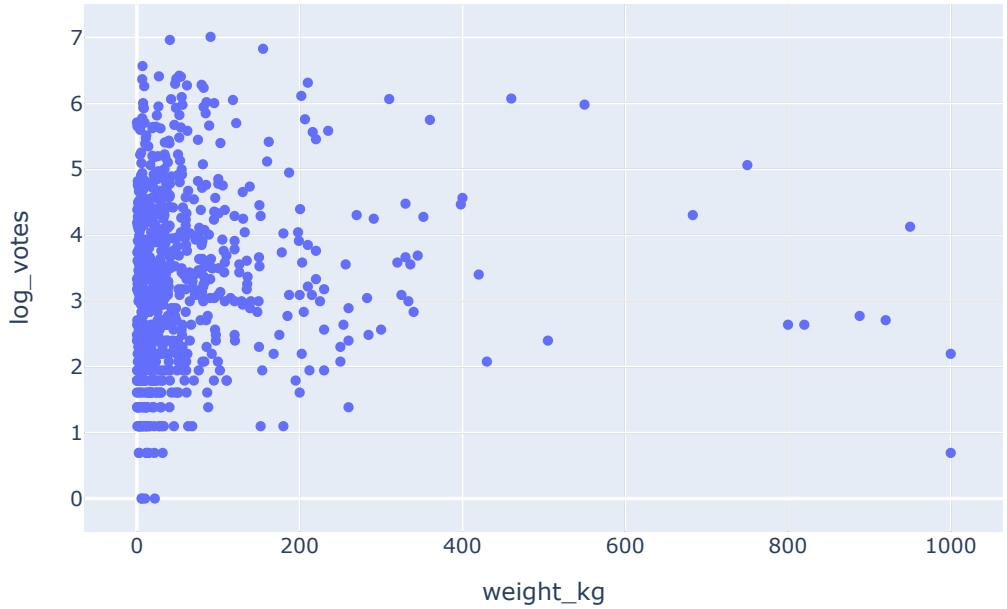
There are some peaks in this data, one around having a base total around 320, a smaller one around 400, and the end of the data does seem to imply an increase. The peak around 320 is the group of very popular cute pokemon, like Bulbasaur and Eevee. The peak around 400 is the group of middle evolutions of popular pokemon. The largest peak at the end are just the big strong most popular pokemon. There does seem to be a trend of increase in base total being related to an increase in votes, although it's pretty sketchy.



It looks like it's not that legendary pokemon are unpopular, but rather since there are much less legendary pokemon than regular ones, and the super popular pokemon are all not legendary so that's pulling the coefficient to say that being legendary has a negative effect.

For height and weight, since Poisson regression assumes a linear relationship between the log of the response and the predictors, plotting the log of Votes vs height and weight we get this:





We can see that basically there's a huge clump of points for small weight and heights, while there are a

few pokemon that have huge weight and height that's messing up the regression. Basically it's a bad idea to use height and weight as factors for prediction.

Conclusions

There is a pretty clear bias for older Pokemon among older fans on Reddit at least, and I hope my poster conveys that. I tried to show that the main reason the older pokemon were liked was exactly because they were older, which was the results of the regression and the specific example.

General Considerations

I don't think the regression model I used in the poster is the best, cause I don't know how to account for Generation probably being a confounding factor. The effects of the specific pokemon types are probably not correct cause generation is a confounding factor. Plus it could just be that the variation is decreasing as generation increase, although the median also decreases as generation increases, so I think there is some merit in my claim that popularity of pokemon are decreasing as generation increases.

I feel like using the reddit popularity poll results as a measure of how well liked a Pokemon is may have some issues. Mainly, the reddit survey made each voter specify 1 Pokemon as their absolute favourite, although a person may like a wide variety of Pokemon, and so a Pokemon with very little votes may not actually be a Pokemon very disliked, but rather a Pokemon that most people are just neutral about. As well, if a person's favourite pokemon is a gen 1 pokemon I don't have any access to their preferences about other pokemon, unfortunately. If I also had the voter's first pokemon game played that may also assist in making my point that people tend to like pokemon from their first game played, due to nostalgia.