Data Brawl

The personality predictor

In this project we will investigate the data set of people answering a questionnaire about their personality type, whether they are introverts or extroverts.

Introduction: understanding the data

Table 1: Data description

To understand what we have in hand, an initial analysis of the data set is needed. To accomplish that a table 1 is presented with summaries of the dataset and their main categorical data. Firstly, note that in order to accomplish this some of the data has had to be cleaned, therefore those who had unreasonable values for the datatype have been cleared (eg age of 250), as well as some individuals who refused to answer some of the main questions about themselves. As the dataset is so large, dropping these 147 observations doesn't represent a big loss for the study, only a mere 2% decrease in sample size.

	extro	intro	neither	Overall
	(N=965)	(N=4339)	(N=1737)	(N=7041)
Gender				
female	595 (61.7%)	2493 (57.5%)	1055 (60.7%)	4143 (58.8%)
male	352 (36.5%)	1722 (39.7%)	625 (36.0%)	2699 (38.3%)
other	18 (1.9%)	124 (2.9%)	57 (3.3%)	199 (2.8%)
Country				
Austria	55 (5.7%)	187 (4.3%)	73 (4.2%)	315 (4.5%)
Canada	69 (7.2%)	269 (6.2%)	100 (5.8%)	438 (6.2%)
Germany	24 (2.5%)	125 (2.9%)	41 (2.4%)	190 (2.7%)
Great Britain	67 (6.9%)	276 (6.4%)	159 (9.2%)	502 (7.1%)
Indonesia	19 (2.0%)	64 (1.5%)	24 (1.4%)	107 (1.5%)
Other	238 (24.7%)	1356 (31.3%)	539 (31.0%)	2133 (30.3%)
United States	493 (51.1%)	2062 (47.5%)	801 (46.1%)	3356 (47.7%)
Age				
Mean (SD)	26.5 (11.8)	26.1 (11.3)	25.5 (11.4)	26.0 (11.4)
Median [Min, Max]	22.0 [14.0, 72.0]	22.0 [14.0, 81.0]	21.0 [14.0, 90.0]	22.0 [14.0, 90.0]
Native in English				
No	256 (26.5%)	1407 (32.4%)	509 (29.3%)	2172 (30.8%)
Yes	709 (73.5%)	2932 (67.6%)	1228 (70.7%)	4869 (69.2%)

Above the general description of the data is clear in that between extroverted people and introverted people, there are no great shifts in response rate, other than that of gender, where it can be identified that females tend to identify more as extroverted relative to men. Other than that all the other categorical variables seem pretty evenly split between categories, including the group which doesn't identify as either extrovert or introvert.

Responses of the survey:

Turning now to the survey data, some of the data of the questionnaires shows interesting patterns worth looking into. For instance, the relationship of the time spent in each question and the order of the questionnaire is quite important, as it may be depicting errors, or people who lost engagement throughout. Also, the relationship in the pattern of answering is relevant as well, such as whether people tend to answer more in binary packages such as (1-5) or more of a gradient, using the values 2-3-4 rather than the extremes.

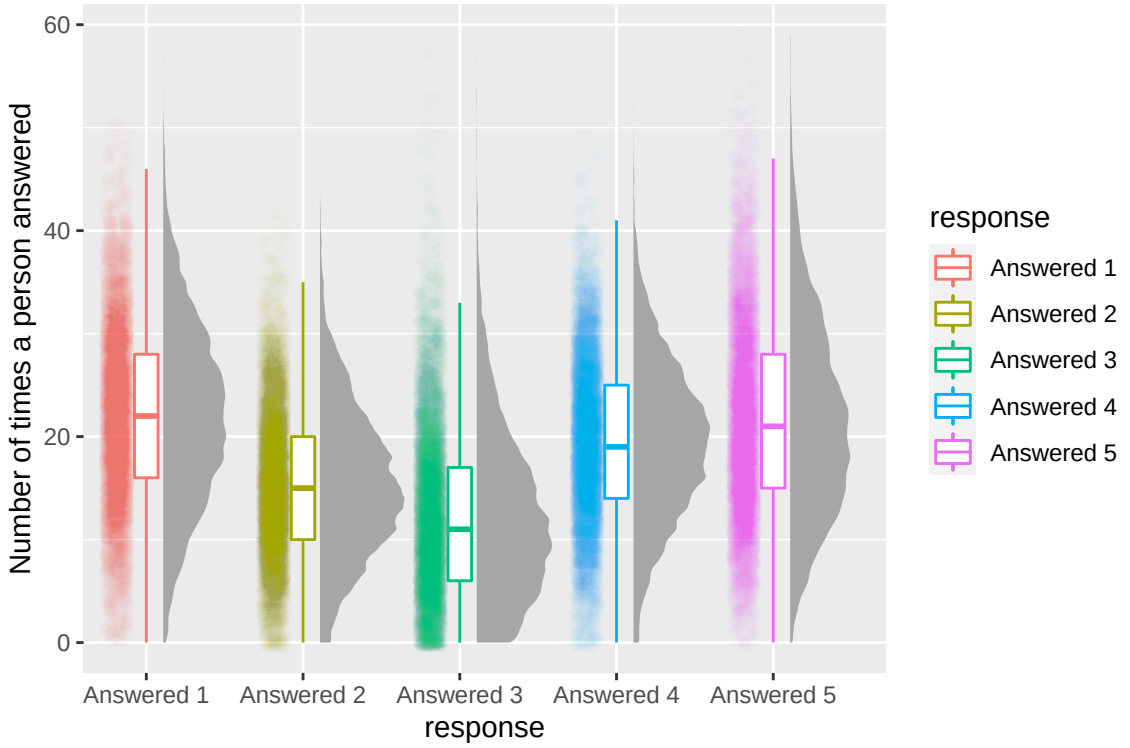
Time patterns:

In order to understand time patterns, filtrate all the responses in which look unreasonable, for instance if people spent more than 5 minutes in a question they will be filtered away as that will not be representative for this particular study. Furthermore, in the calculation of the regression in how the order of the questions it's important to also take into account the extra time that non-english speakers spend in each question when answering.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3424563	0.0230789	231.48618	0
positions	-0.0093751	0.0003498	-26.80116	0
english_nativeYes	-0.7041322	0.0198937	-35.39474	0

In this case, the results display that on average people spent 5 seconds in each question, but progressively lost interest and answered faster and faster, with response time falling 9 milliseconds at each extra question. Furthermore, if the individual was a native English speaker, the response time was 0.7 seconds faster per question. This results signal don't show any preoccupying features, as for the last question people will take on average 4.5 seconds, around 15% less than at the start of the survey.

Now looking into the answering patters of individuals, whether people answer all the same or people display different "styles" of answering. Below is a box-plot displaying how many of each score (1-5) was used by each individual person. Clearly the extremes, numbers 1 and 5, were the favorite choices, with people choosing to score 4 on the questions as a close second. Also, the distributions look nicely shapped normally, so there aren't any further issues that should be investigated.



Now moving more closely onto the trade-offs made by people, it's interesting to investigate whether there is a response style where people choose "gradient" answers and a different group which chooses only "extremes". In order to look into this, plot the counts for each score and investigate relationships.

In the plots above it can be seen how there is an inverse relationship between the extreme people and the undecided people, who choose more gradients and midpoints rather than extremes. Moreover, further analysis yielded as well that there were no significant differences between categorical groups (introverts/extroverts, males/females etc.) at the time of answering, so it's determined by unknown variables.

The Analysis:

Now that the data has been outlined and the main issues resolved, the next step is to look into the answers and what they mean in terms of our research questions, are you an introvert or an extrovert?

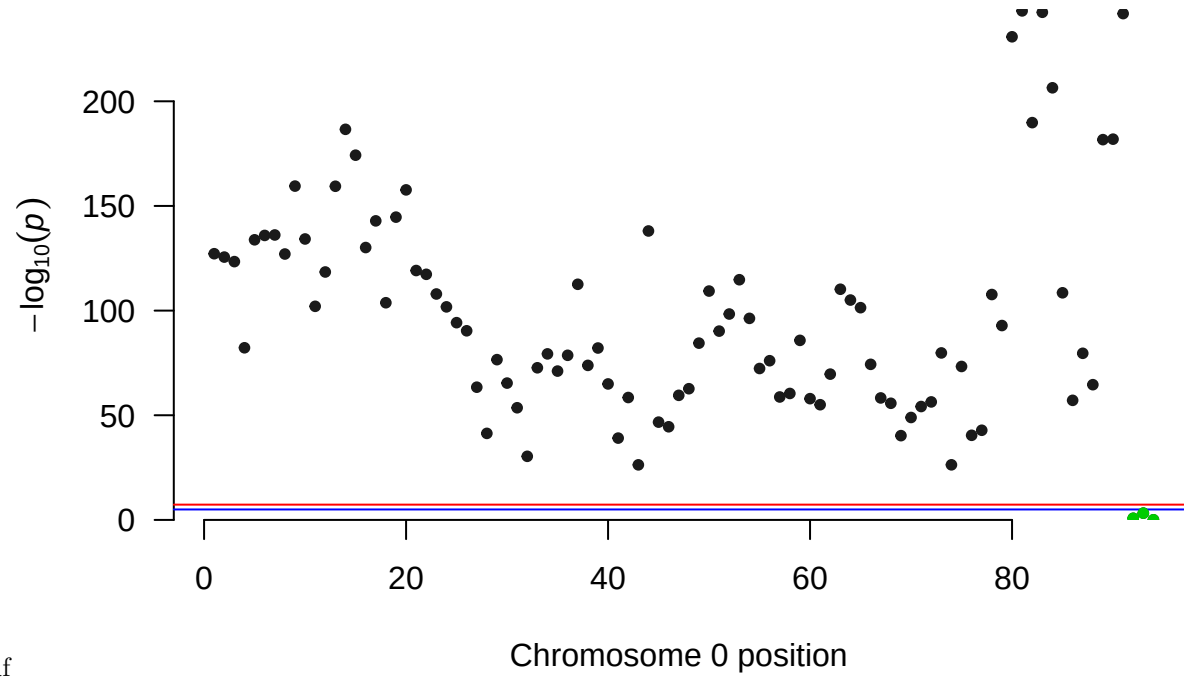
Borrowing from Genetics...

The first analysis way will be to borrow an analytical technique used in genetics to try and identify the most relevant questions in the analysis. As such, it requires considering each question individually and whether is any good at determining whether people are introverts or extroverts. Using logistic regression, consider the following:

$$\text{Odds of being Extroverted} = exp(\beta_0 + \beta_1(\text{Points given to Question }_i)\ )$$

The meaning behind this equation suggests that, excluding the baseline level accounted for in β_0, how much will an extra point of the scale (1-5) awarded to a question, make it more likely that the individual ends up answering they are introverted/extroverted. However, in this first part it's not important the *magnitude of how likely they are to be either*, but rather it's whether the question is meaningful or not. To evaluate this, the *p-value* of the equation is taken, and if they are sufficiently small, then consider (Points given to Question $_i$ as an interesting question to the study, else it will be "useless".

The technique shown below is a Manhattan plot showing the size of the p-values for each question.

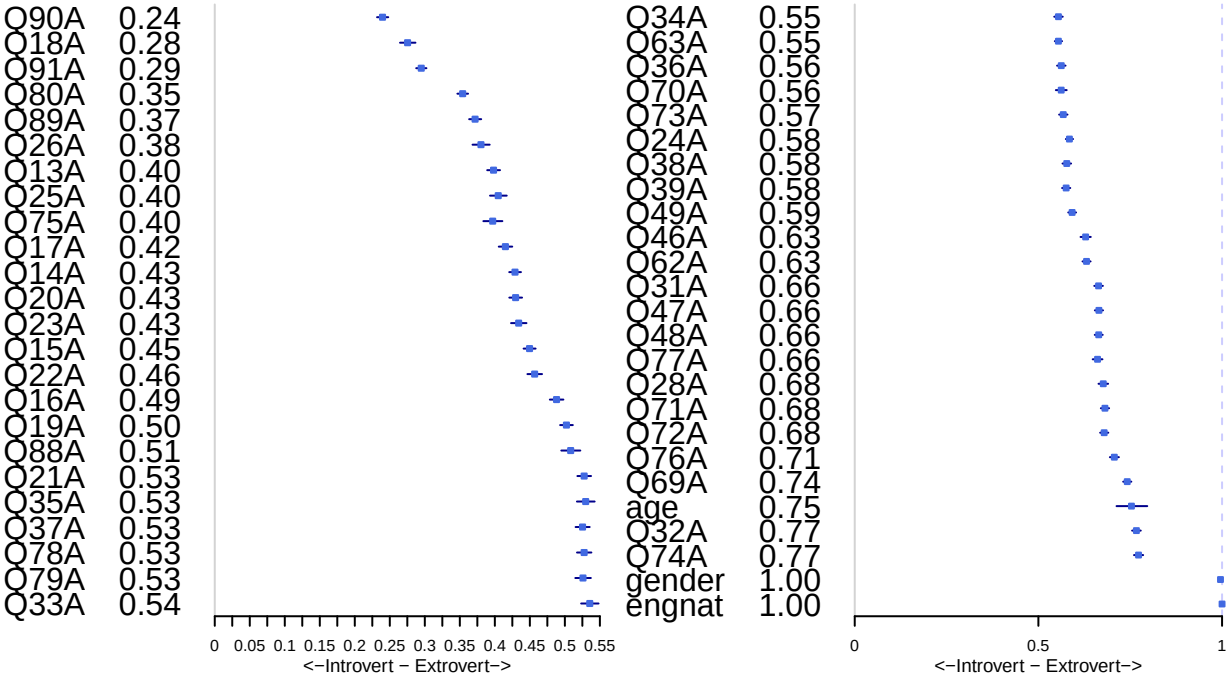


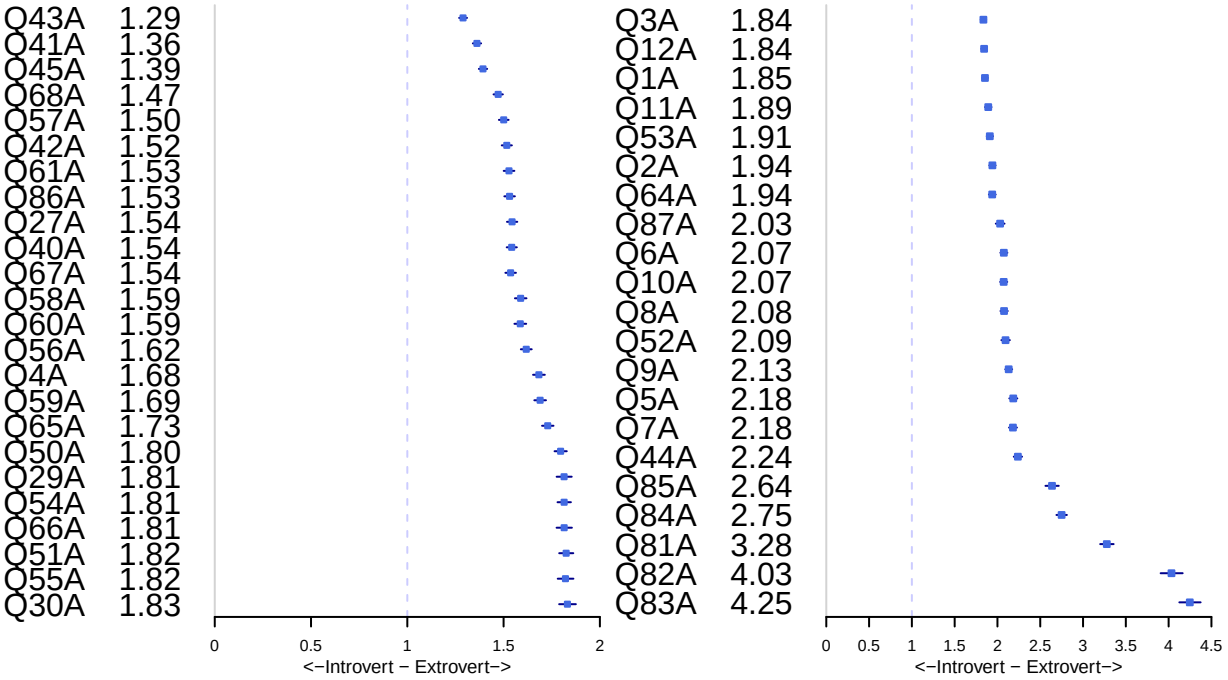
finish-1.pdf

The results are quite astonishing, in the plot clearly shows the questionnaire is an extremely good predictor when it comes to nailing whether the people responding are or not extroverts/introverts, at least with a highly significant outcome (shall see on the size of the effect later). For instance, in green it can be seen how dummy variables, such as native language, age of the individual or the time they spent on the test, all showing clearly no significant relationship on whether they are or not extro/introverts.

Magnitude of effects:

Having thus determined it's worth looking at all the questions in the questionnaire, it's time to proceed to compare the size of the effects and which ones are better predictors. In order to get a general idea of the direction in which each questions points (ie whether people are introverts/extroverts) and by how much, a forest plot is used to layout all the effects. The plot below tells you what is the odds of being an extrovert, (ie \$1-Odds = \textrm{% chance of being extrovert by every extra point you answered in that question} \$. For example, if the odds is 0.5, then you have a -50% chance of being an extrovert for every extra point you selected in that question: or equivalently the higher you scored that question, more likely you are an introvert. Similarly, 1.5 (ie 50% chance) indicates if you scored 5 you are 5x50% more likely to be an extrovert compared to someone who selected 0. (note 0 is impossible but the regression doesn't know that...)



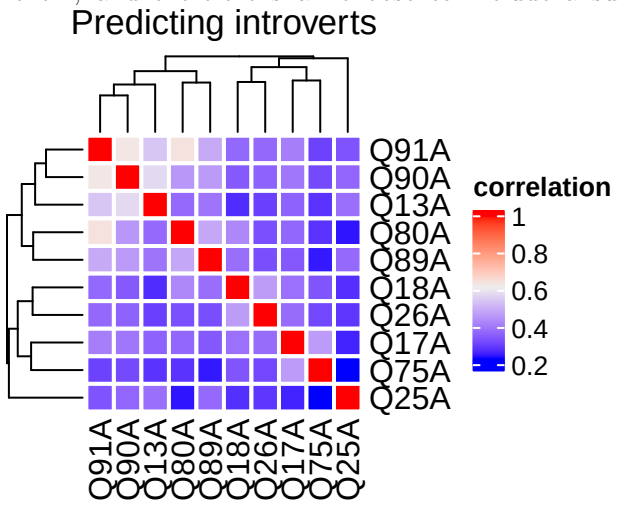


Anyhow, the results are extremely satisfying with the output being a great deal of variety, ie some questions leaning extrovert and other introvert, with smaller and greater magnitude of effects.

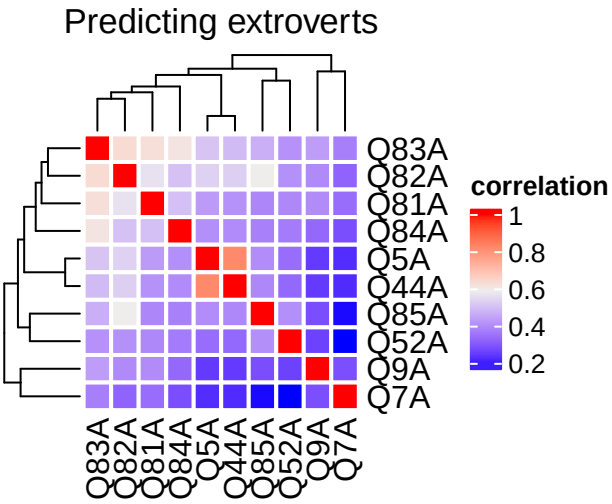
Model building exercise

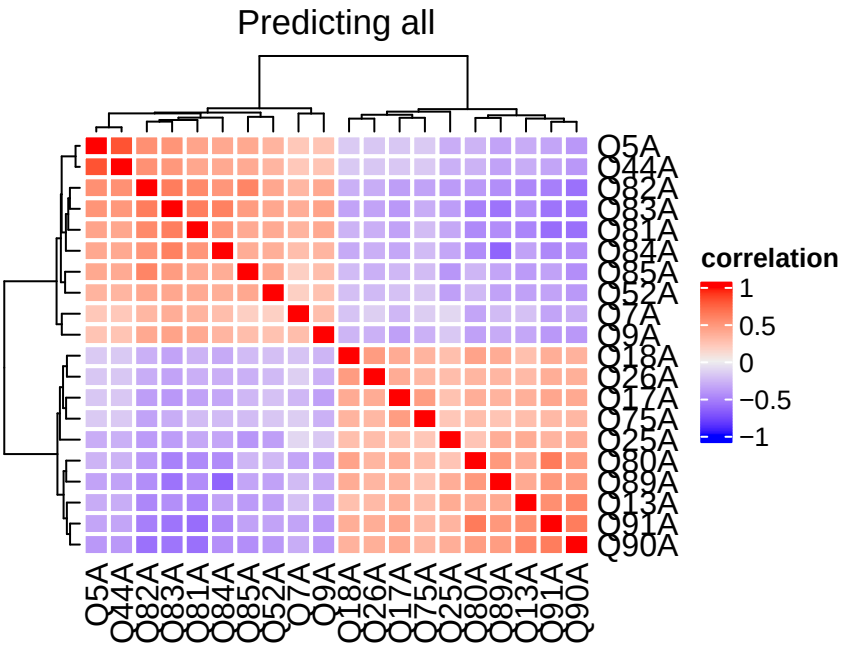
Now that the magnitude has been displayed, it would be too messy to use all questions in order to predict an outcome, multicolinearity comes into play and the results wouldn't be ideal and, as always, the simpler the better. Therefore, filter through the top and bottom questions of those that better predict our outcome.

First step is to look at how correlated those variables are, as if they show high correlation of outcomes, it's dubious we can use them, and therefore shall choose to include a subset rather than all of them. - top bottom-1.pdf



- top bottom-2.pdf





The results of the correlation matrix show that while some of the variables are slightly correlation, said issue is only happening in the first plot, only happens between two of the questions (Q5 - Q44), all the rest seem quite correlation free.

Now move into using this information to build our predictive model, first looking into the introvert-deciding questions:

	Est in %	Lower CI	Upper CI	p-value
(Intercept)	331914.64	264733.14	416138.39	0.000
Q83	-40.12	-42.47	-37.66	0.000
Q82	-43.50	-45.86	-41.04	0.000
Q81	-38.12	-40.13	-36.04	0.000
Q84	-31.55	-33.76	-29.27	0.000
Q44	-10.57	-13.62	-7.42	0.030
Q85	-8.68	-12.50	-4.68	0.153
Q7	-37.82	-39.78	-35.80	0.000
Q9	-25.39	-27.41	-23.32	0.000
Q6	-9.27	-11.97	-6.48	0.030

Building the model, one of the variables has been removed (Question 5), as per the previously identified high correlation with other variables. Now doing the same with the variable identifying the extrovert variables:

	Est in %	Lower CI	Upper CI	p-value
(Intercept)	-100.00	-100.00	-100.00	0.000
Q17	24.46	20.22	28.85	0.000
Q25	27.57	22.57	32.78	0.000
Q13	23.40	19.31	27.63	0.000
Q75	16.82	11.61	22.27	0.022
Q26	14.32	9.55	19.28	0.034
Q89	43.09	38.72	47.59	0.000
Q80	58.93	54.09	63.91	0.000
Q18	34.32	27.45	41.56	0.000
Q90	87.14	79.44	95.18	0.000
Q91	50.18	45.13	55.41	0.000

This time since the correlation matrix didn't show any issues, the results remain untouched. Thus move to build the final model.

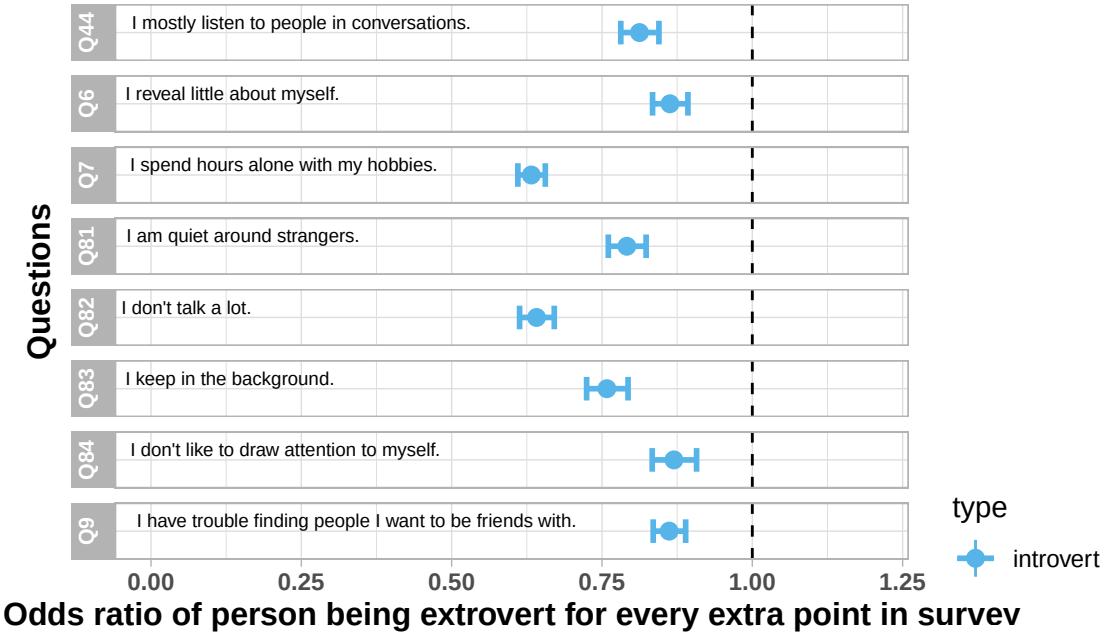
When comparing the whole model, it gets more complex, since there will be interactions between the questions and it needs to be checked using other methods, such as comparing which is the AIC criteria (best model decider). As a result, Questions 85, 75 are further dropped, and using a nested model check called a Chi-square test, the results yield that the model including all 20 questions isn't any better than the model with only the 17 identified. The model output can be seen below, with the coefficients being, as discussed above % odds of being extroverted.

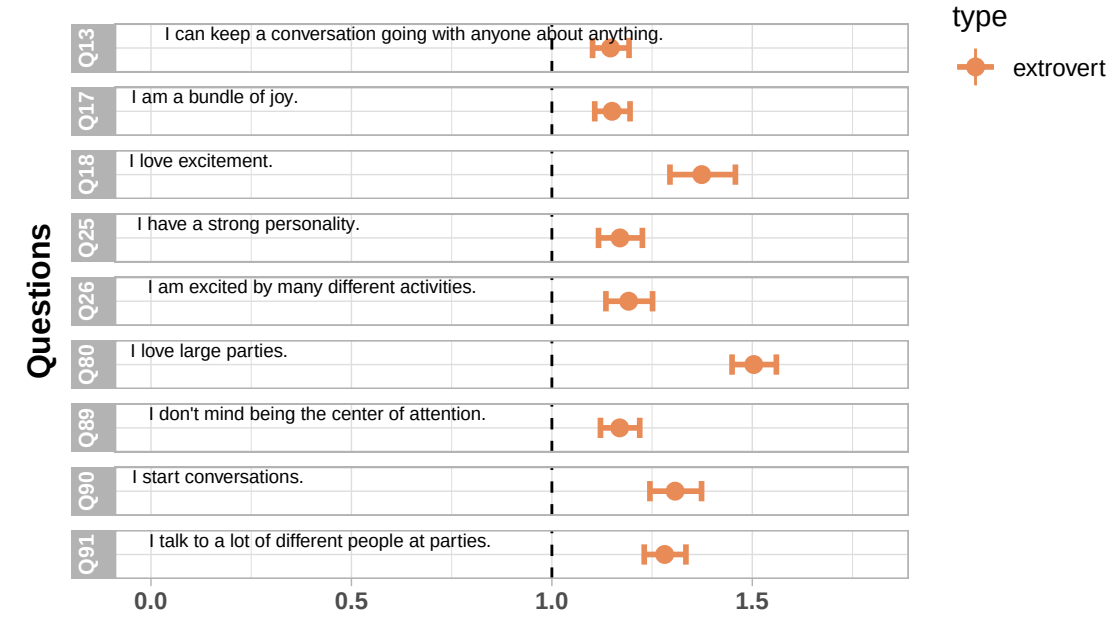
	Est in %	Lower CI	Upper CI	p-value
(Intercept)	-80.87	-88.09	-69.26	0.019
Q17	15.01	10.70	19.50	0.014
Q25	16.98	11.64	22.58	0.024
Q13	14.61	10.16	19.25	0.020
Q26	19.16	13.48	25.12	0.015
Q89	16.92	12.12	21.91	0.012
Q80	50.37	44.95	56.00	0.000
Q18	37.37	29.45	45.78	0.000
Q90	30.74	24.45	37.34	0.000
Q91	28.13	23.04	33.44	0.000
Q83	-24.19	-27.55	-20.66	0.000
Q82	-35.88	-38.70	-32.93	0.000
Q81	-20.87	-23.95	-17.66	0.000
Q84	-13.04	-16.65	-9.27	0.026
Q44	-18.76	-21.87	-15.53	0.000
Q7	-36.75	-38.99	-34.42	0.000
Q9	-13.83	-16.48	-11.08	0.001
Q6	-13.70	-16.59	-10.71	0.004

In the results, therefore we can see Q7 and Q80 being the strongest effect predictors (50.37 towards introvert and 35.88 towards extrovert) respectively. All the rest of the variables also have a significant effect and therefore should definitely be kept in the dataset for exploration. Below is the breakdown of exactly which questions are revealed to be the most influential in our model (note it's in untranformed % ie 0.5 is 50% decrease while 1.5 is 50% increase in odds of being extrovert):

```
## Help on topic 'sort' was found in the following packages:
##
##   Package               Library
##   base                  /Library/Frameworks/R.framework/Resources/library
```

```
##    BiocGenerics           /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##
##
## Using the first match ...
```





Final model checks and play:

Finally, all that's left is cross validate the results and see what percentage of our sample is well predicted using these calculations. In order to do this, first it's going to estimate what percentage of the sample is well fit using the fixed parameters estimated. After this has been calculated, the more flexible cross validation method will be used, iterating through 1/10 of the dataset and estimating the model using the remaining 9/10, thus seeing whether it would be a good model to estimate foreign data with.

```r
cost.misc <- function(response, pred) mean(abs(response-pred) > 0.5)
cv.err1 <- cv.glm(manh_database, b_t_10_mod_2, cost.misc,K=10)$delta[1]#

cv.err3<- cv.glm(manh_database, b_t_10_mod_2, cost.misc,K=500)$delta[1]

# %  people which correctly predicted?
1-cost.misc(as.numeric(manh_database$IE)-1,fitted(b_t_10_mod_2))
```

```
## [1] 0.9368253
```

```r
#%  people which correctly predicted?
1-cv.err1
```

```
## [1] 0.9362255
```

```r
1-cv.err3
```

```
## [1] 0.9362255
```

The results of doing this validation, yield that the validity of our data is 93%, meaning that it's possible to predict accurately 93% of the people who enter the questionnaire. Equally, using the more advanced validation method also yields a similar result, which could be a feature of the dataset as it's relatively uniform. In the output above, the first is the one using the fixed results, and the second are using a partition of 10 and 500 parts of the dataset respectively.

And now prediction:

Finally, we will try to predict the outcome of those who identify as neither extroverts or introverts: Using the formula, and the code below, the final conclusion can be taken that 37% of those in the "neither" category are actually introverts and the rest are extroverts, at least using our predictive model.

```r
#Proportion of introverts:
b/length(a)
```

```
## [1] 0.0005757052
```