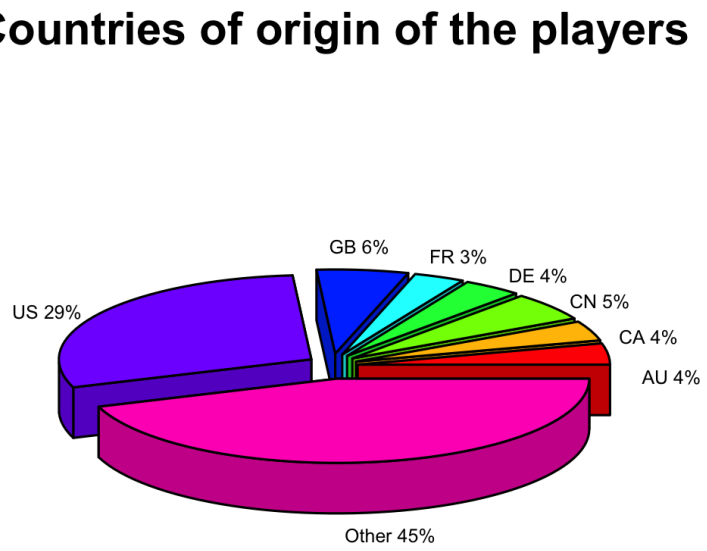
Data-Scientist Test Data

A video-game company created this small database to test candidates who apply to a data-scientist position. I wrote a R script to access their data through SQL queries (using both the RSQLite and sqldf packages).

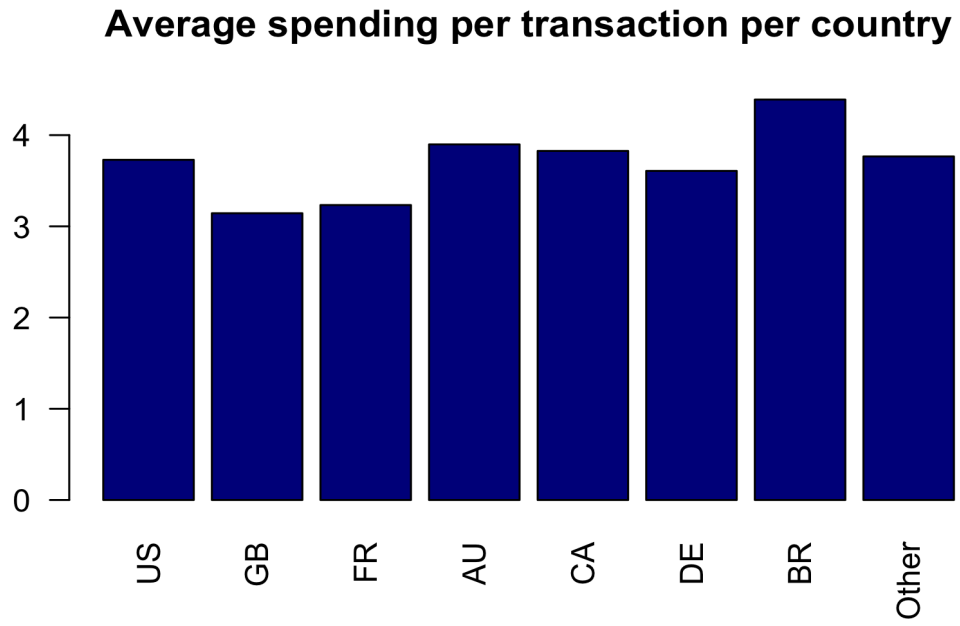
Here are the answers to the test questions (units to quantities like the cash amount are not provided in the test data, so none are shown here):

```
Total revenues for 2013/02/01: 159.64
Number of users using two different device:  0
The country producing the most revenues is: US
The iPad/iPhone split in Canada is: 205 iPads, and: 328 iPhones
The proportion of lifetime revenue generated during the first week
is: 74.40579 percent
```

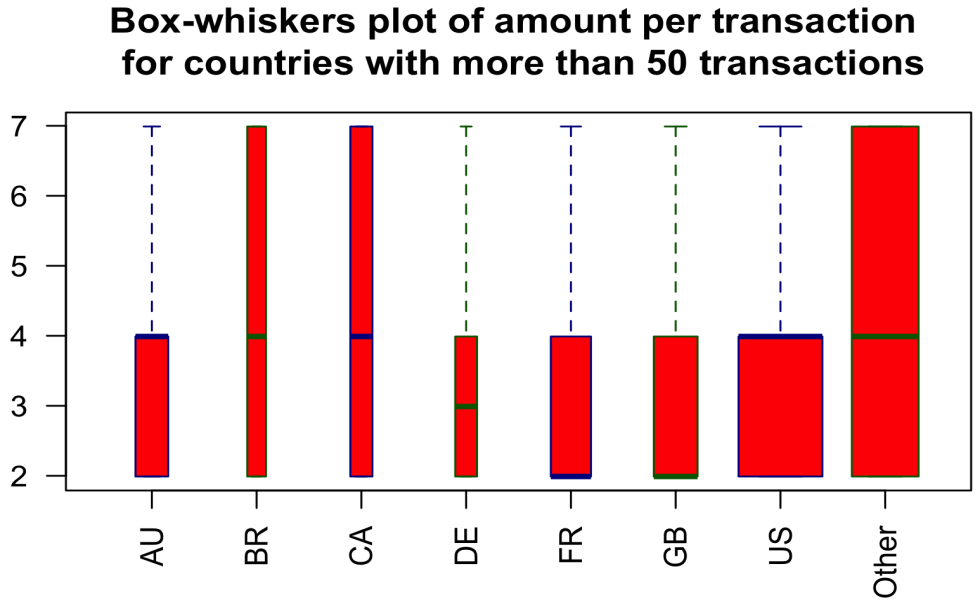
Now for the visualization part, here are different plots to emphasize some aspect of the data:
First, the repartition by countries of the players. We only show the first 7 countries (in terms of numbers), and list all of the others under the "other" label.



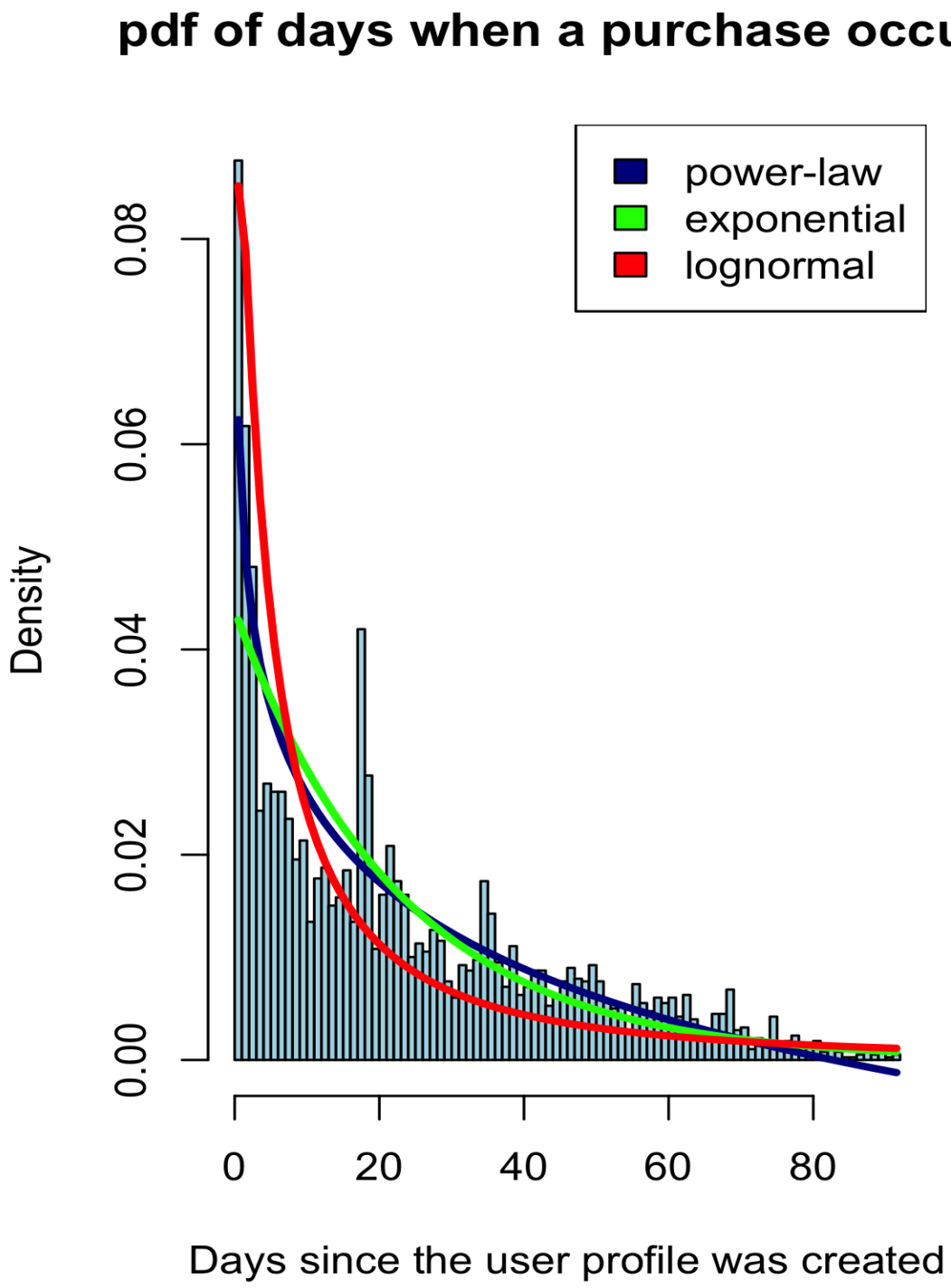
Unsurprisingly, the US is their biggest market in number of users. Then let's look at the average revenues per country per transaction (we only show the 7 countries with the most transactions. You will notice that those are not exactly the same as the countries with the most players, previously shown):



All of these countries are pretty similar in terms of average spending. BR (Brazil?) seems to bring the more cash per transaction though.
Let's study the distribution of spending per country, again for the countries with the most transactions. A box-and-whiskers plot can help visualize basic information about such distributions. In this case though, it is not very useful as the prices fall in only three categories: 1.99, 3.99, and 6.99. Still, it

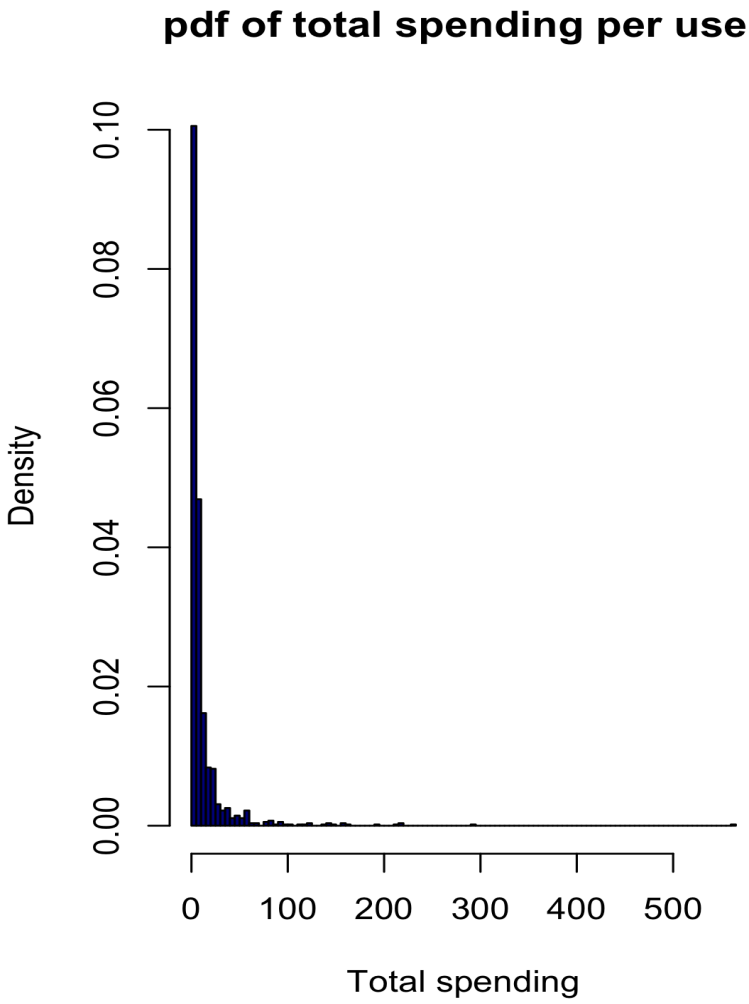


shows that in the country with the most transactions (the US) the median revenue per transaction is 3.99, and that's also the 3rd quartile (75% of the players spend 3.99 or less per transaction). This 3rd quartile is lower than for the "other" category. The width of the boxes is proportional to the number of transaction per country. You will notice that the distributions appear strongly skewed.
Next, we can study the distribution of days when an in-app purchase is made (from the time a user created his/her account). Here is the probability density function:



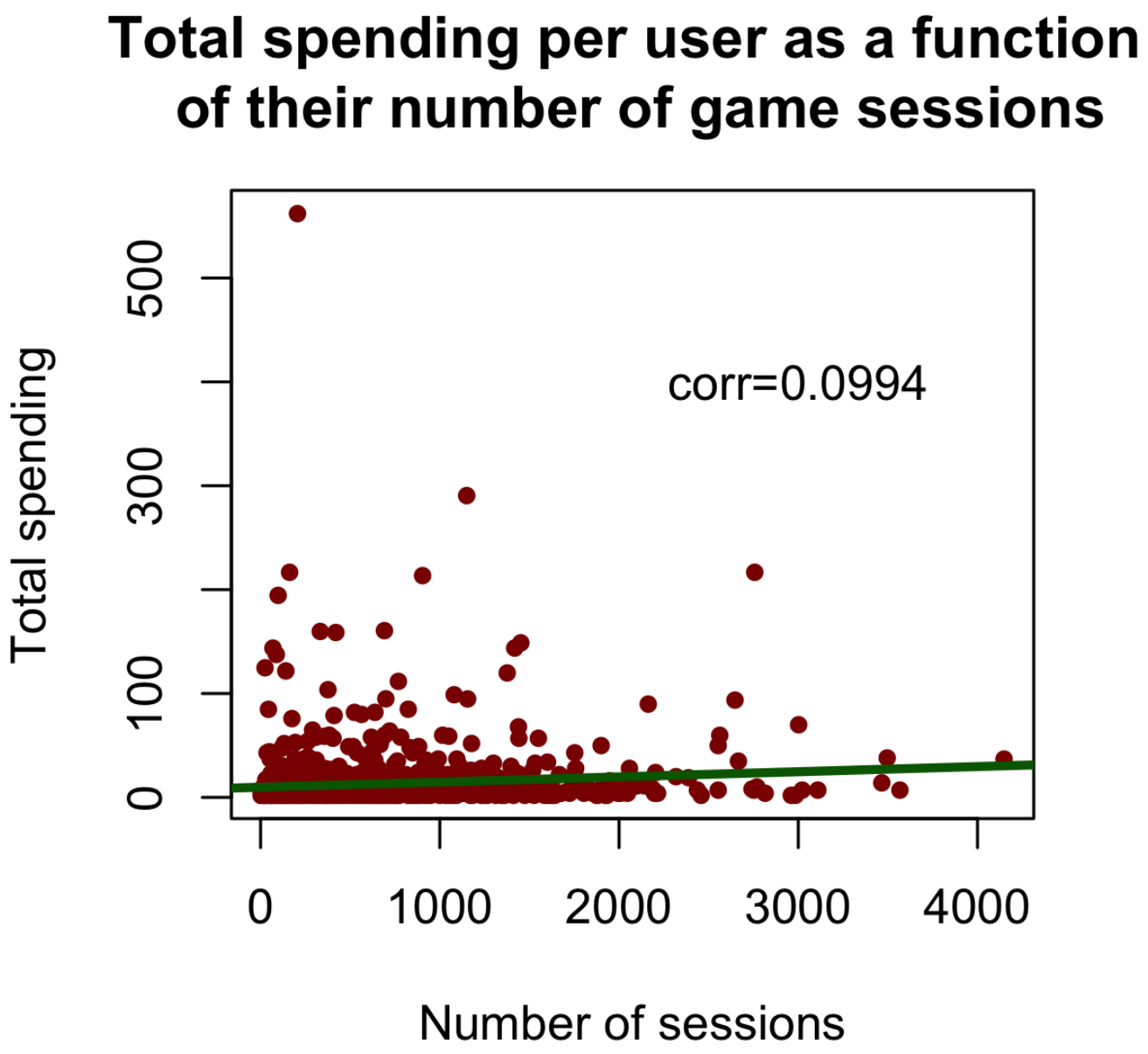
This pdf was fitted by three different distributions: power-law, exponential, and lognormal. A basic "sum of squares" argument to decide which distribution is the best fit returns the power law. However, the lognormal seems to do a better job at fitting the tail of the distribution. This has consequences about what the company should do with its customers: the lognormal distribution implies that a player can still make in-app purchases more than 80 days after this player started to play. Therefore, the company should keep engaging such players (even if, clearly, the vast majority of purchases are made the first days after the player created is/her account).

We can also look at the distribution of total spending per user:



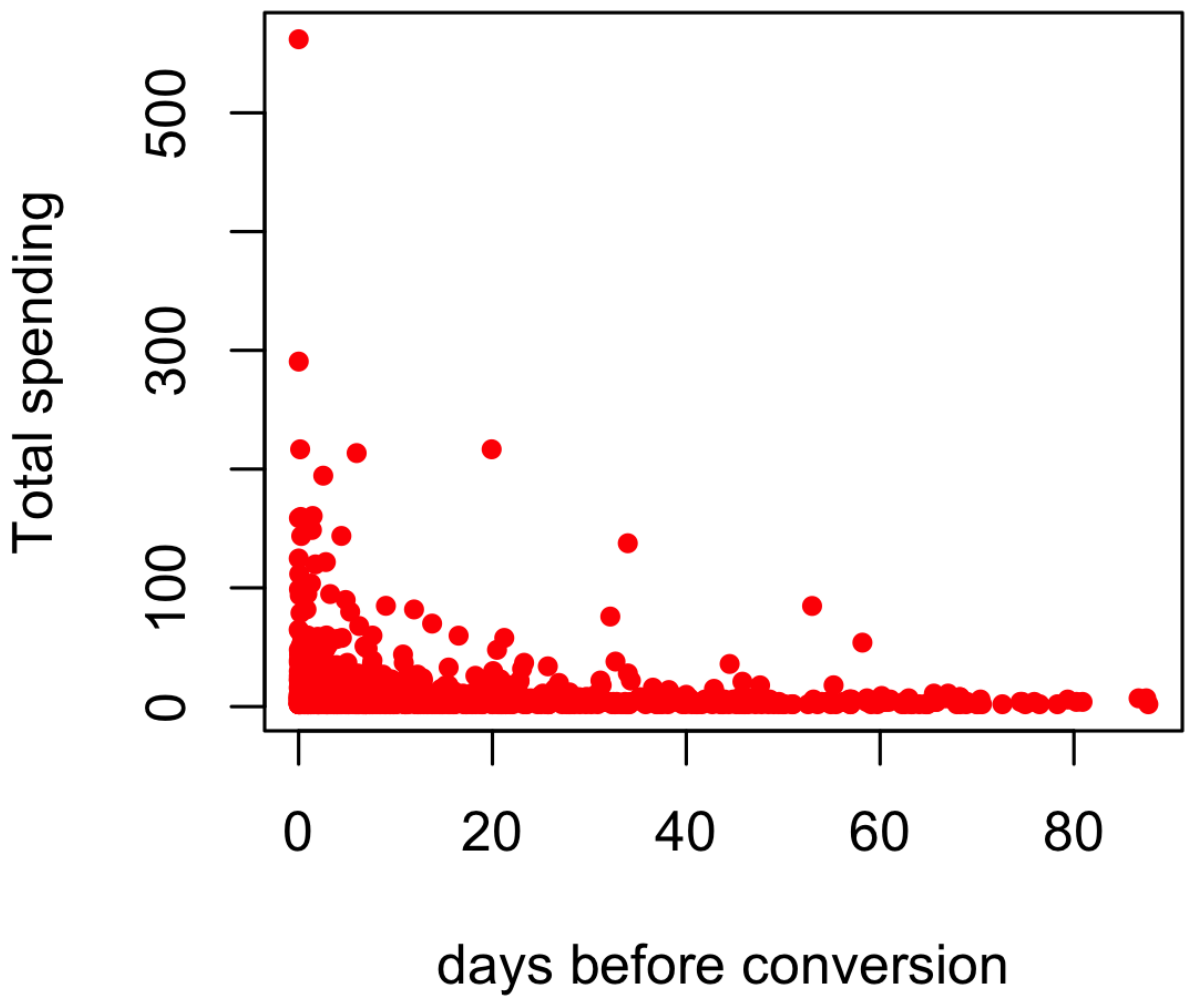
The probability density function shows that the vast majority of users only spend a small amount of money on the game, but you have a small numbers of users who spend significantly more (you can notice a non-zero probability of spending more than 500). The company probably does everything it can to retain such high-value players.

Can we find some relations between the total spending per user and some useful predictor variable? Here we look at the total spending as a function of the number of sessions a user plays. The idea behind it is that the more sessions a user plays, the more engaged he/she is with the game. Is he/she then more likely to spend more on this game? The following figure shows that this is not the case.



The green line is the result of a linear regression, but as you can see the fit is not good.
In fact, the Pearson correlation coefficient is very weak, at only 0.0994.
Therefore, it is safe to conclude that the number of sessions a user plays the game is not a good predictor of the amount of money he/she is going to spend.

Here is another weak relation, between the total spending and the number of days to conversion. Conversion refers to the fact that a user who played the game for free suddenly accepts to make in-app purchases. The following plot shows that the relationship between the number it takes a player to convert and the total amount of money he/she is going to spend on the game is really weak:



There seems to be an inverse relationship: the faster a user convert, the more money he/she is going to spend on the game. However, the Pearson correlation is only -0.136 so the relation is very weak (of course, since the relation between the two variables is visibly not linear, the Pearson coefficient is not very useful. The Spearman rank coefficient returns -0.237, which also confirms a very weak negative correlation. Because the p-values for both correlations are smaller than a 5% significance level, then the null hypothesis that these correlations are actually 0 can be rejected).

The data-scientist test requests a machine learning algorithm. Here, I did a **logistic regression** using two predictor variables (again, the days before conversion and the number of game sessions). Of course, since I already know that these variables are not good predictor, the logistic regression won't perform very well.

Indeed, the result of:

```
summary( glm(spending ~ res$delay + res2$max_sessions,
family="binomial"))
```

is:

```
Call:
glm(formula = spending ~ res$delay + res2$max_sessions, family =
"binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8751  -0.3404  -0.3016  -0.2066   3.3774

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.977794   0.224739 -13.250  < 2e-16 ***
res$delay         -0.049838   0.015056  -3.310 0.000933 ***
res2$max_sessions  0.000673   0.000196   3.433 0.000596 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 418.89  on 1099  degrees of freedom
Residual deviance: 394.78  on 1097  degrees of freedom
AIC: 400.78

Number of Fisher Scoring iterations: 7
```

For the dependent variable, which has to be categorical (0 or 1), I took a total spending larger than 50 (variable equal to 1) or lower than 50 (variable equal to 0).
The p-values in this logistic regression are very low for both predictors, confirming their uselessness at forecasting the total spending of a player.

I could have tested other predictor variables (like the age of the player, which I believe shows potential as older players may have more disposable income than younger ones. Unfortunately, the `birth_year` field of the table `account` seems to be only filled with `NA`.