Analysis of router setup data

Gustavo Fernandez

10/30/2020

Abstract

Motivation: The purpose of this analysis is to demonstrate how to pull signal from the noise to draw insights.

Data: Data_analyst_assessment.csv contains a sampling of data from a hypothetical router's setups. At each screen of the setup process, a row is created in this data_analyst_assessment table.

Research Question: The main questions to answer in this analysis are:

1) where are customers struggling in the setup process? What should my team focus on first?

2) What is average elapsed time of incomplete setups (ones that are missing a setup complete event).

Method: Even though test like anova can be run to compare means across groups, that was not the main focus of this analysis. Visual methods, data description, data cleaning and manipulation were employed and were found to be suitable to answer the research questions.

Software: Most of the data exploration, manipulation and visualization were performed using R software, libraries such as ggplot, dplyr, and Amelia were employed. SQL queries were executed inside SAS software.

Exploratory Data Analysis

The data set contains 5721 observations and 8 variables. The list of variables and general structure of the dataset is shown in table 1.

Table 1. Variables in the Data Analyst Assessment Data Set

Name	Description	Type	Missing (%)
Setup_id	each time a router is setup, all events tied to that setup are given this unique ID	Factor	0%
Platform	The platform through which the router was setup	Factor	52%
Type	a category for the event	Factor	0%
Timestamp	the time of the event	Numeric	0%
model		factor	0%
step	the specific page in a multi-page setup the	Factor	2.9%
duration	elapsed time of the setup process at that event	Numeric	12.29%
Rating	a self-reported 1-5-star rating of the setup process	Factor	98.5%

Like a lot of data, this dataset is messy, with some inconsistencies and some strange characteristics. 21% of the values in the data are missing. (figure 1).

As shown in table 1 the variables with the most missing values are rating (98.5%), follow by platform (52%). There is also an important amount of missing values in duration (12.29%) and in the category step (2.9%).

A map of the missing values distribution, in the data set can be seen in figure1.

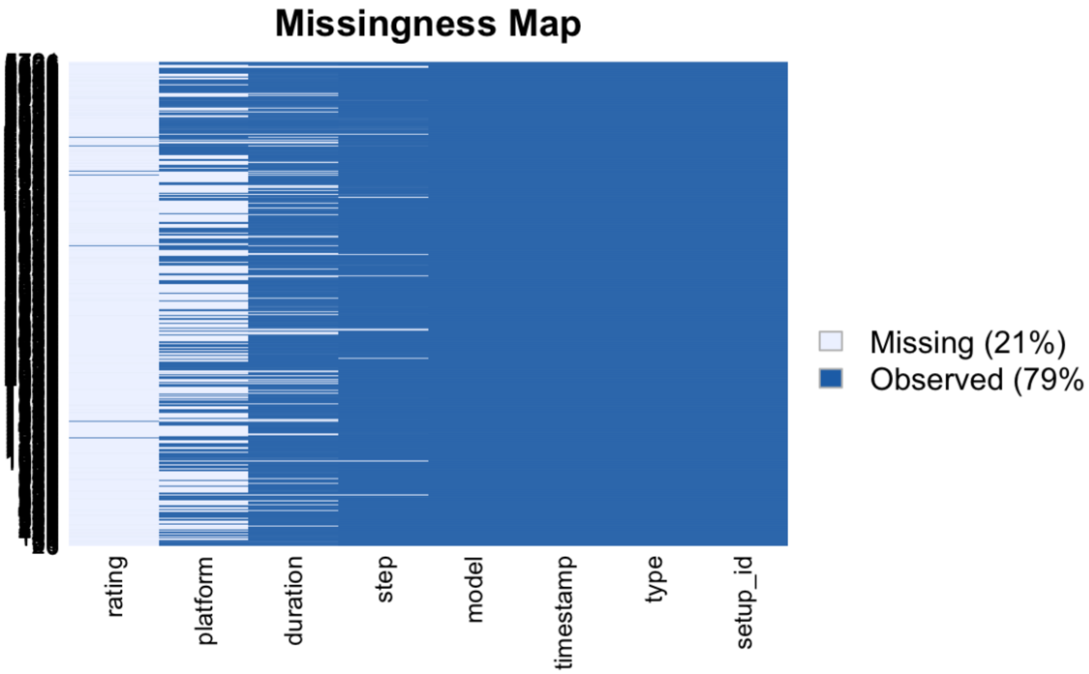


Figure 1. missing values in the Data Analyst Assessment Data Set

Due to the high amount of missing values of the variable rating and since I consider this variable not an important measure to describe the situation or help in answering the main questions of this analysis, I decided to remove it and no longer consider it for further analysis. Likewise, the variable model was removed. On the other hand, despite the high amount of missing values of the variable Platform I consider it to be important since the success or failure on completing a setup might be associated with the type of platform in some way or another. As well as the possibility that it might be a difference in duration in one or many of the steps between the two platform. For these reasons, I decided to not remove it and consider the missing values of this category as "not specified". Moreover, the missing values in the variable step were replace by the input ("others"). I decided to remove the observations with missing values of duration. As well as those with a value greater than 1 E+6 since I consider them to be non-realistic. Furthermore, deleted the observation setup_id 80 since it was making the dataset messier by adding a level on the variable type by a single observation. The final data set for the analysis contains 4,996 observations with 453 distinct setup Ids. The list of factors and their levels are shown in table 2.

The number of times each step appears in the data set is visually shown in figure 2 below.

Table 2. Levels of the factors

Name	Levels
Setup_id	453 distinct Ids
Platform	3 levels (iOS, Android, not specified)
Type	3 levels (setup error, setup rating, setup step)
Step	28 levels

Distribution of step

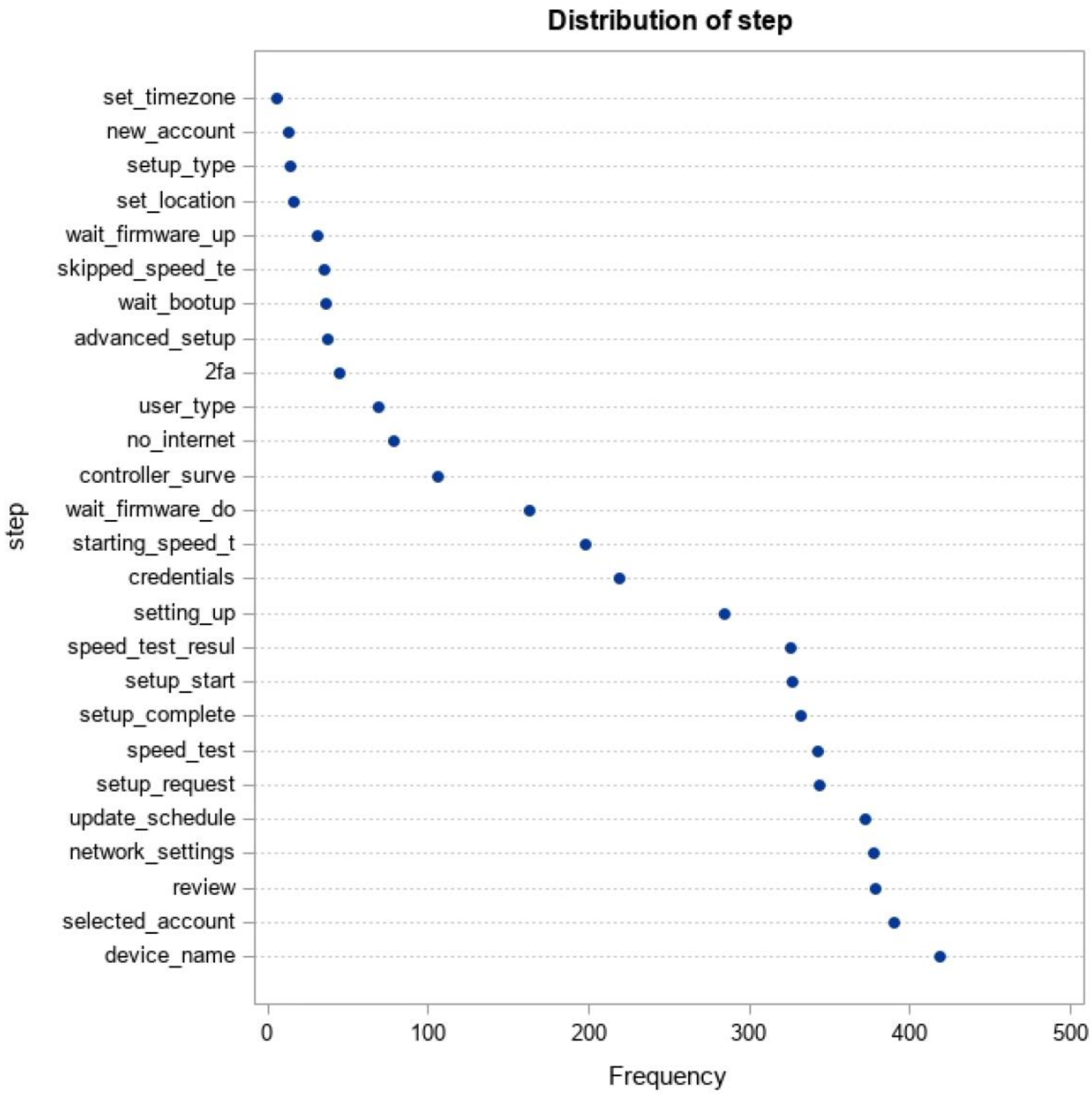


Figure 2. frequency of step in the data

Duration by Step

Controller survey, advanced setup, setup complete and set location had the highest median duration which might indicate that in those steps the costumers are struggling more.

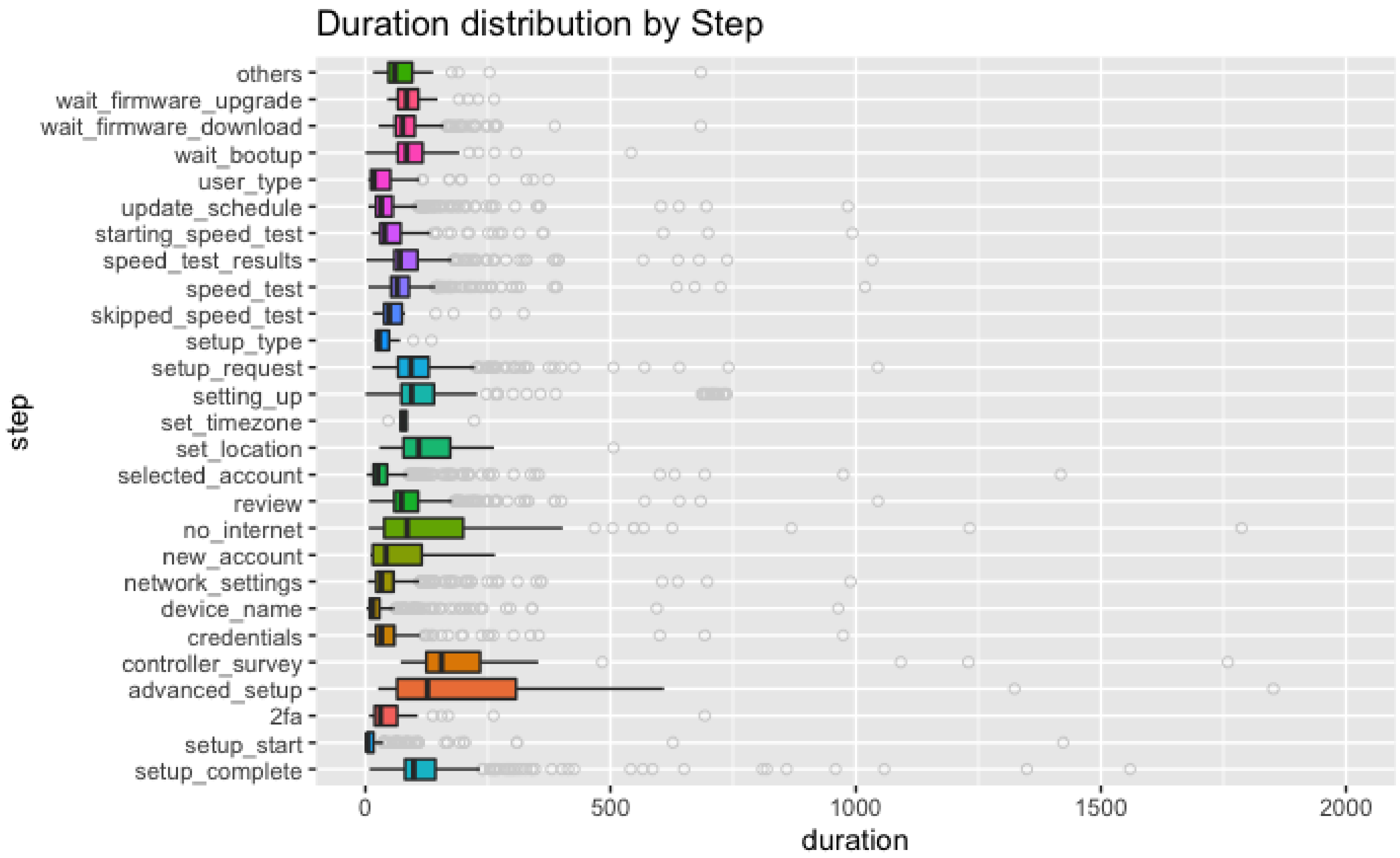


Figure 3. Duration distribution by step

These steps also show high variability specially advanced setup and no internet steps which indicate huge potential for improvement if we can identify the process or features affecting the variability in the duration of those steps.

The corresponding median and IQR of the duration of the most relevant step are shown in table 3.

Table 3. Median and IQR of the most relevant steps

Step	Median (duration)	IQR (duration)
Controller survey	157.93	124.45
advanced setup	127.24	242.67
Setup complete	99.11	62.46
Set location	109.15	95.68
No internet	86.05	163.08

With the purpose of exploring further the variation of these identified steps a plot of the distribution of duration vs steps by platform used is shown in Figured 3. The median set location and IQR are noticeable lower on Androids than on iOS, similar situation can be observed in the controller survey step.

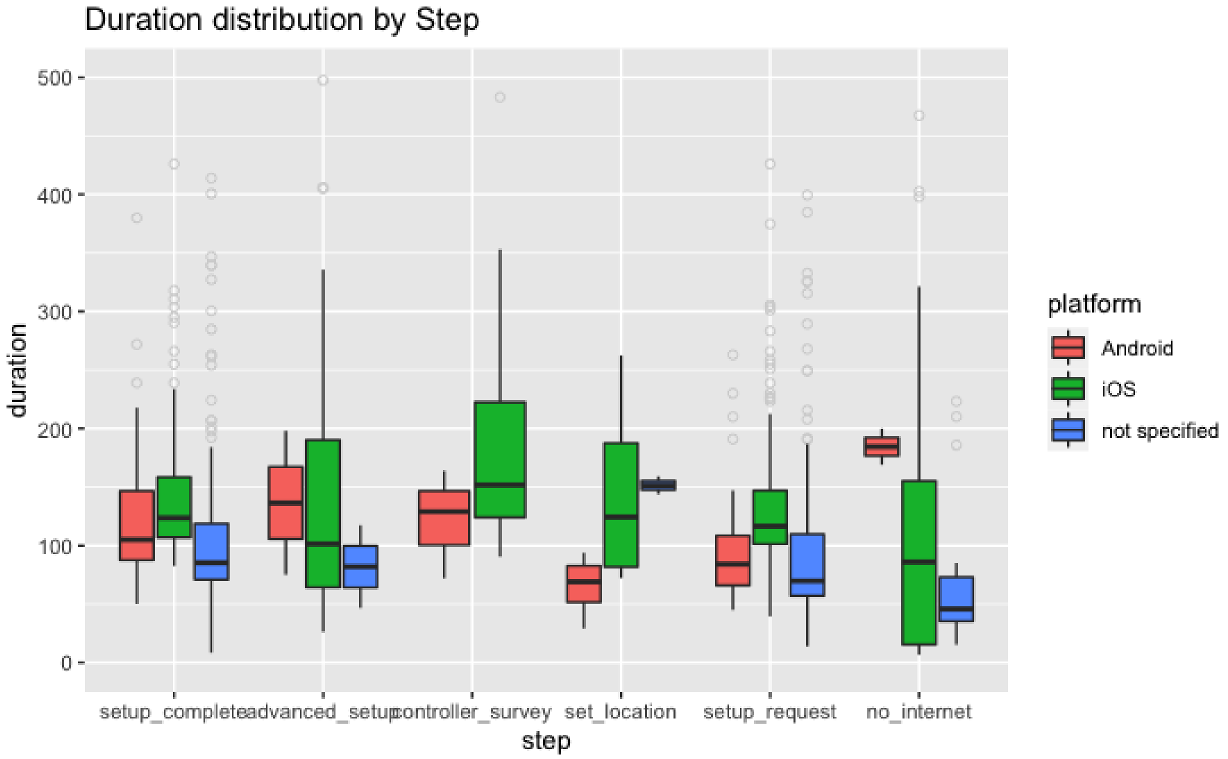


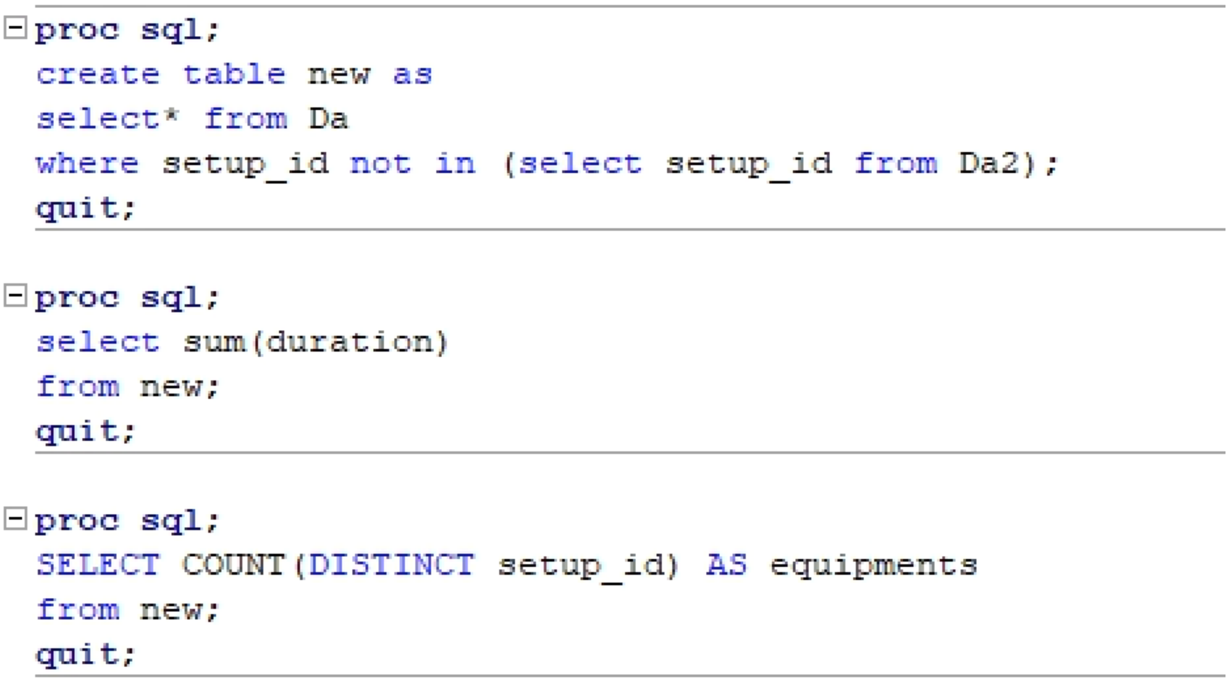
Figure 4. Duration distribution by Step.

The median duration during the no internet step is noticeable higher on Androids, however this is based only on two samples from android. iOS platform displayed extremely high variance in the no internet step. However, a more balanced distribution between Platform is needed since in this step only 2 Androids were shown in the data vs 59 iOS and 17 not specified. It would be interesting to collect more data in order to obtain a more representative sample of the distribution of Android vs iOS to determine if this difference between them are consistent.

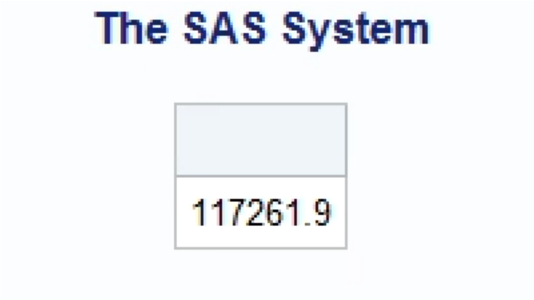
To answer the second question a new table was created by selecting from the original table only the setup_ids that completed the setups, we call that table Da2.

Then, the following SQL queries where executed in SAS to create a table called new which contain the information of only the setup_ids that did not complete the setup. The average elapsed time was calculated by the sum of the durations across the steps divided by the number of unique setups_ids in the new data set.

SQL:

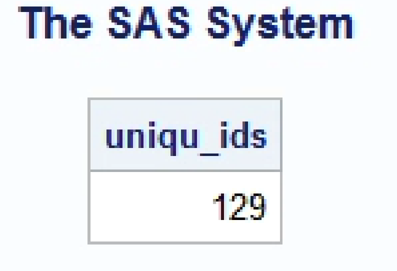


```
proc sql;
create table new as
select* from Da
where setup_id not in (select setup_id from Da2);
quit;

proc sql;
select sum(duration)
from new;
quit;

proc sql;
SELECT COUNT(DISTINCT setup_id) AS equipments
from new;
quit;
```



The SAS System

117261.9



The SAS System

uniqu_ids
129

The average elapsed time of incomplete setups was found to be: 117261.9/129 = 909 seconds.

Conclusion

Controller survey, advanced setup, setup complete and set location had the highest median duration which might indicate that in those steps the costumers are struggling more. These steps also showed high variability specially advanced setup and no internet, which might indicate potential for improvement. If we are able to identify the process or features affecting the variability in their duration.

It would be interesting to collect more data in order to obtain a more representative sample of the distribution of Android vs iOS to determine if the observed difference in the measures of central tendency and spread between them are consistent.

From 453 setup ids in the final data set 129 of them did not complete the setup process (28.48%). The average elapsed time of incomplete setups was 909 seconds. It is important to mention that the analysis and calculations were performed in the presence of outliers. However, these outlier observations were considered to be realistic. For instance, it is possible that a costumer had spent 5,400 seconds (1hour and 30 minutes) performing an advanced setup step.