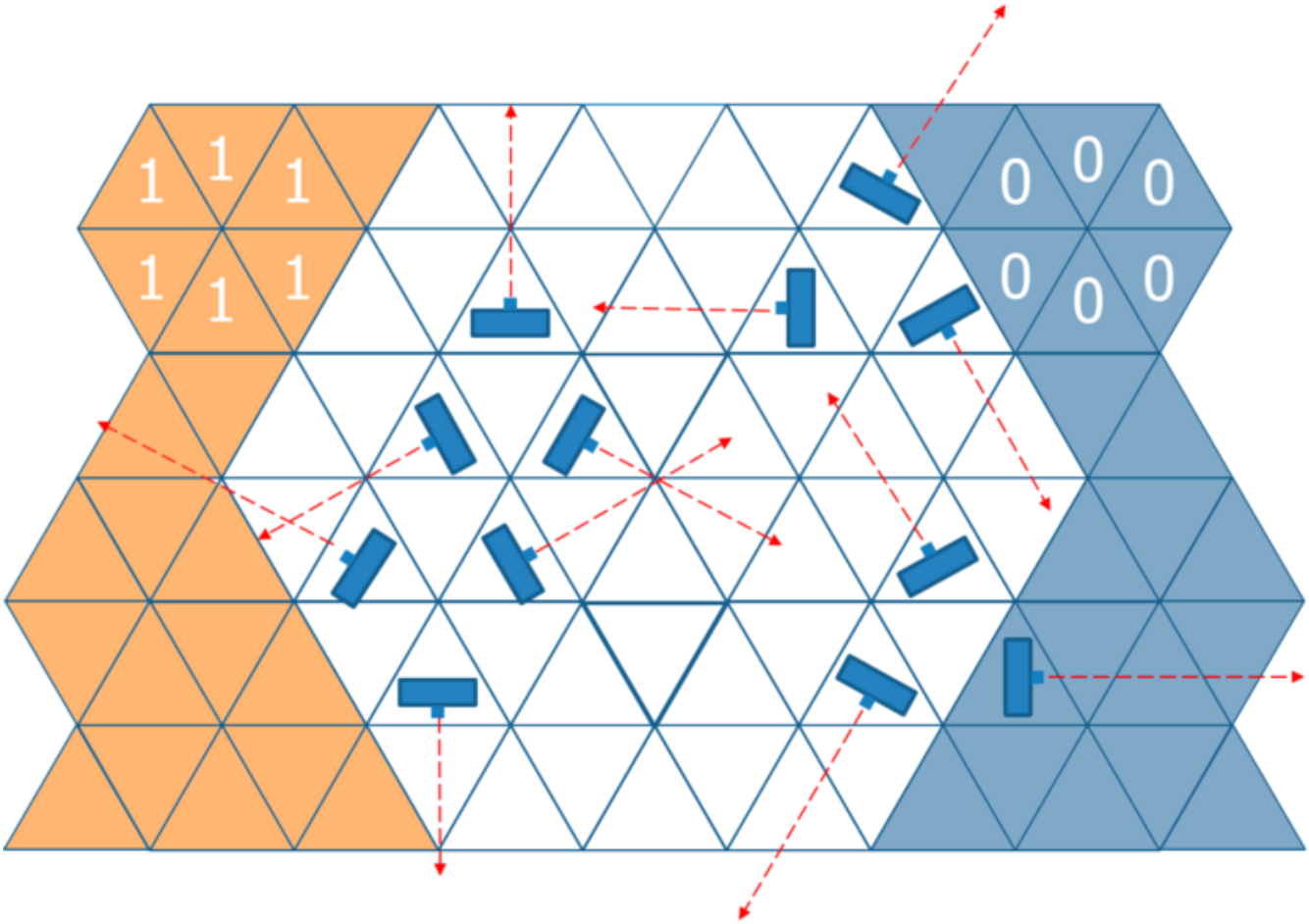


Exploratory Data Analysis with Tableau

Sergey Medvedev
Jun 12 · 5 min read ★

In Machine Learning, an exploratory data analysis or EDA is often the first thing we do to introduce ourselves to a new dataset. It is performed to make general observations about the data, summarize it, explore some basic trends or uncover hidden relations between variables. Data visualisation tools like Qlik or Tableau help better navigate in the new data and present EDA findings in a popular manner. In the next few minutes of your reading time, I will cover a particular classification case with the use of Tableau Prep Builder and Tableau Desktop software.

The dataset we'll be using here isn't new to the town and you have probably come across it before. The data was collected by a Portuguese bank between 2008 and 2013 and contains the results of a telemarketing campaign including customer's response to the bank's offer of a term deposit contract. **Our goal** will be to find those groups of customers in the dataset who are better positioned to react in the affirmative to the campaign. The dataset is available at Irvine's Machine Learning Repository of the University of California. So let's get started!

To begin with, let's connect to the dataset with Tableau Prep Builder, then click 'add step' to learn more about its features — the software will automatically generate for us the summary of all the variables in the dataset:

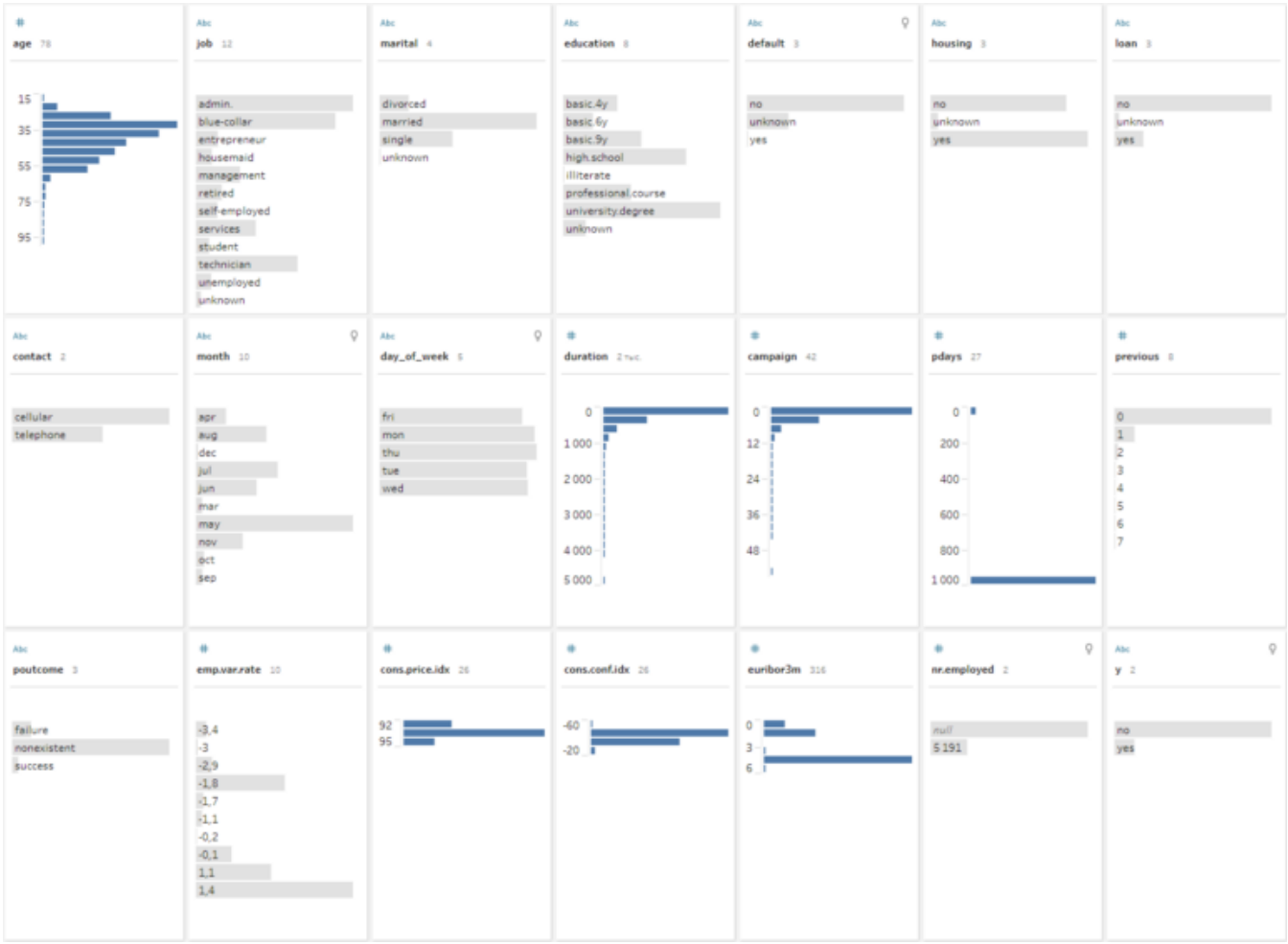
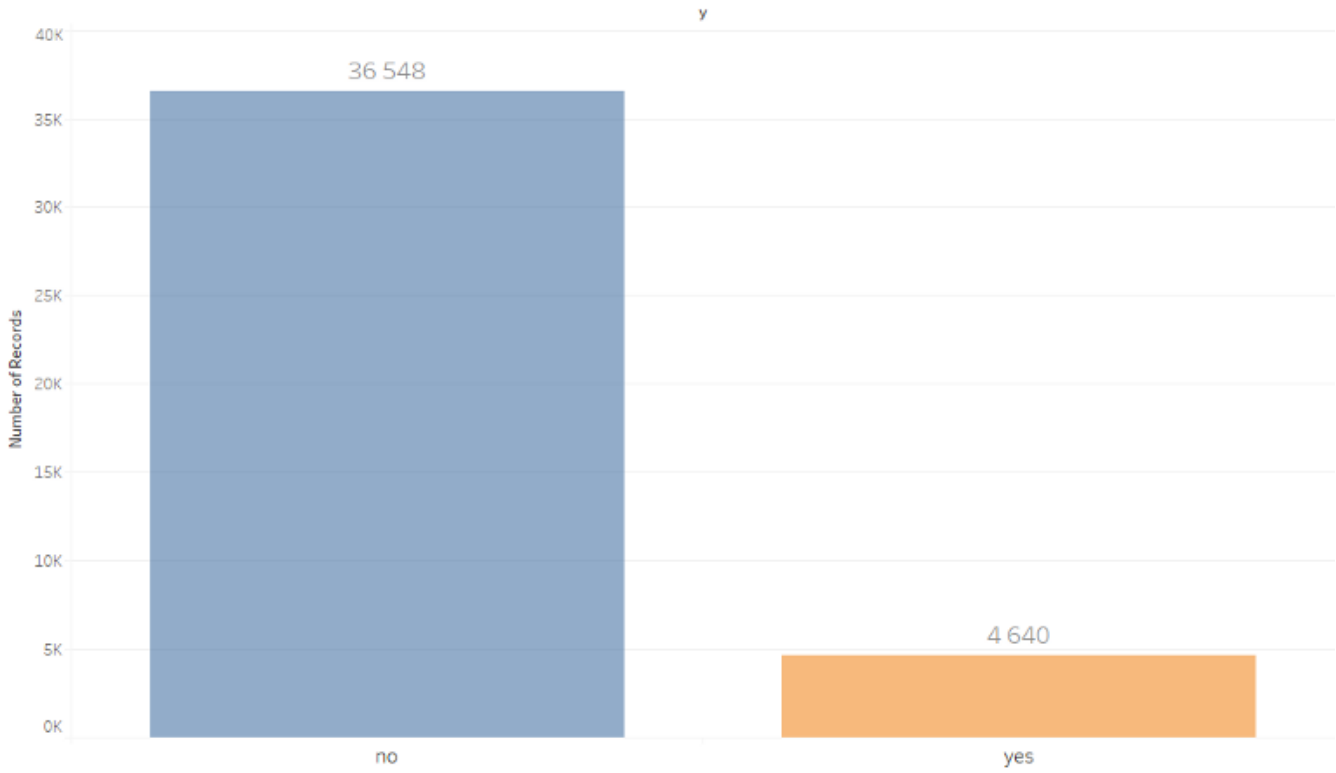


Figure 1. Quick Summary of the Features by Tableau Prep Builder

This is also the point when we use Tableau Prep Builder to introduce some changes into the data, e.g. rename or drop columns, change data type, delete obvious outliers, etc.

Now let's go and explore the data with Tableau Desktop to bring in a few interesting figures. First, let's take a look into the binary target variable 'y' with customer's response to the bank's offer of a deposit contract.

Figure 2. Yes/No Balance in the Target Feature



Class imbalance is the problem that often comes along with such classification cases as fraudulent credit card transactions or the results of online campaigns, for instance. Figure 2 shows that the two classes of the variable 'y' are not represented equally in our dataset also. To be exact, there are 36,548 records belonging to the class 'no' and 4,640 records of the class 'yes'.

The imbalance suggests that later on — if we are going to build a machine learning model with this data — we'll have to oversample or undersample the data prior to training the model on it.

Figure 3. Yes/No Ratio by Age

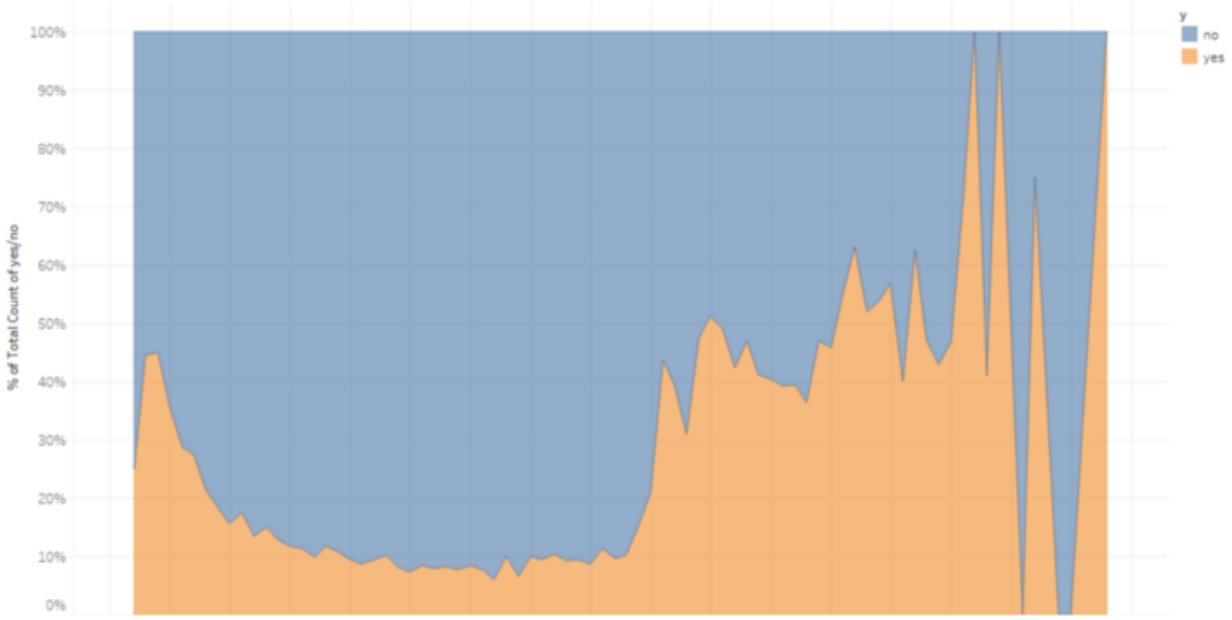
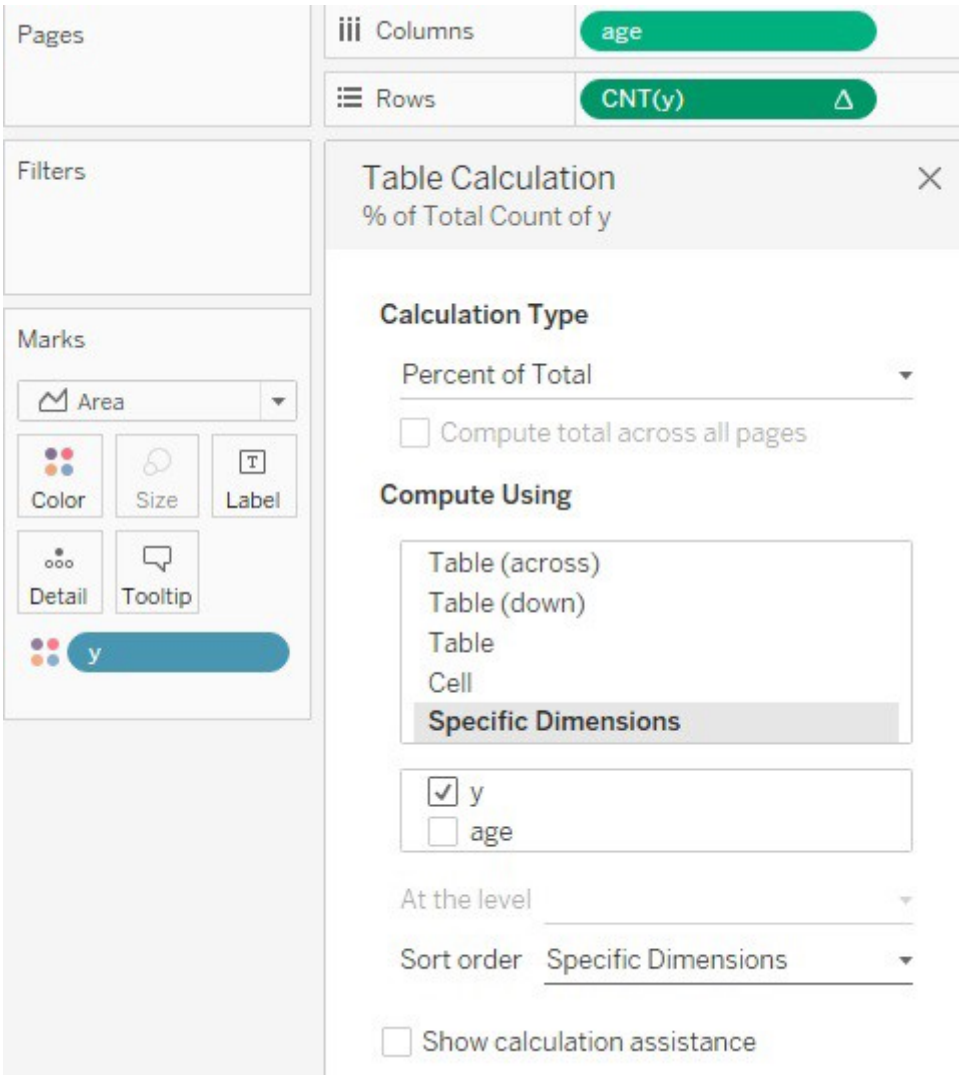


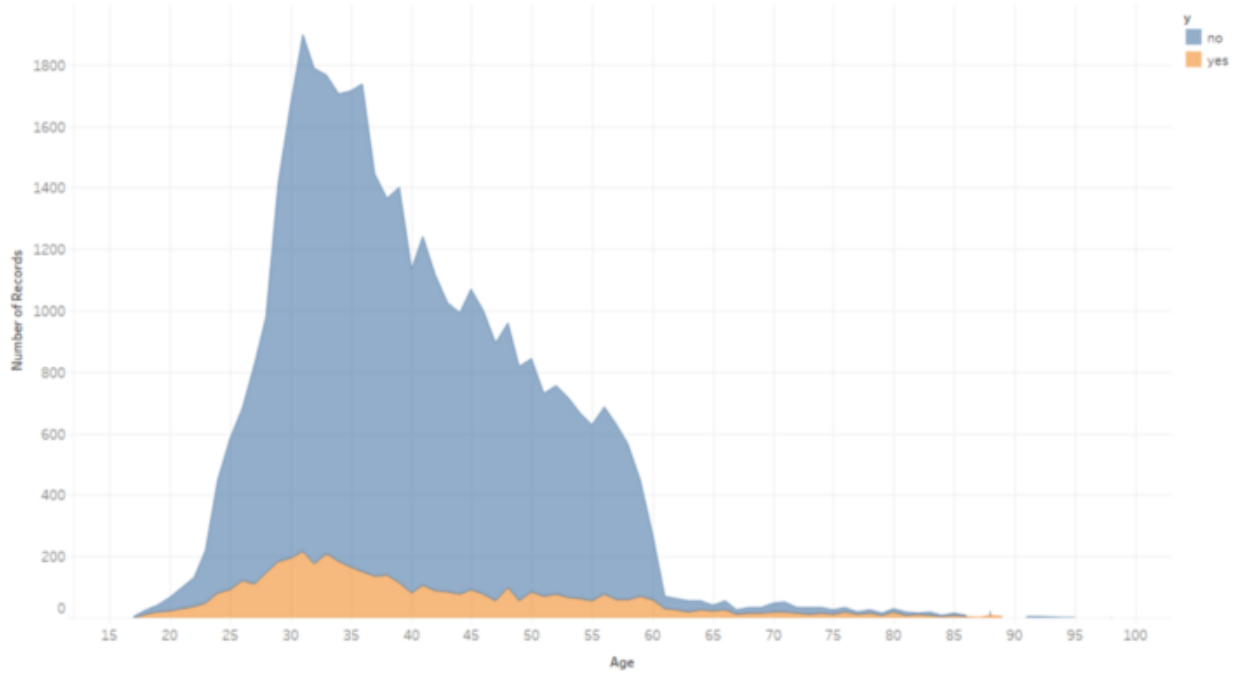
Figure 3 has the ratio between the number of 'yes' and 'no' responses at any given age in the dataset, irrespective of the total number of records. With this graph, we can take stock of how the people of different ages responded to the campaign. From the figure we make a conclusion that the yes/no ratio is better with young people but it also shows a steady negative trend from there on. It reaches a virtual plateau by approximately the age of 30 y/o, and the lowest numbers persist with little change up until approximately a 57 y/o mark. Then the share of 'yes' responses enters the phase of a rapid positive change which might be associated with the usual retirement age. There is simply not enough data in the dataset to adequately represent people over 85 — hence the irregularity of that part of the graph. (To interact with this graph, please click the word Chewbacca)

To create such a figure, you should apply Table Calculation function from the menu of the respected variable (*CNT(y)* in this case), choose Percent of Total as a calculation type and specify 'y' as the dimension for calculating percents, like this:



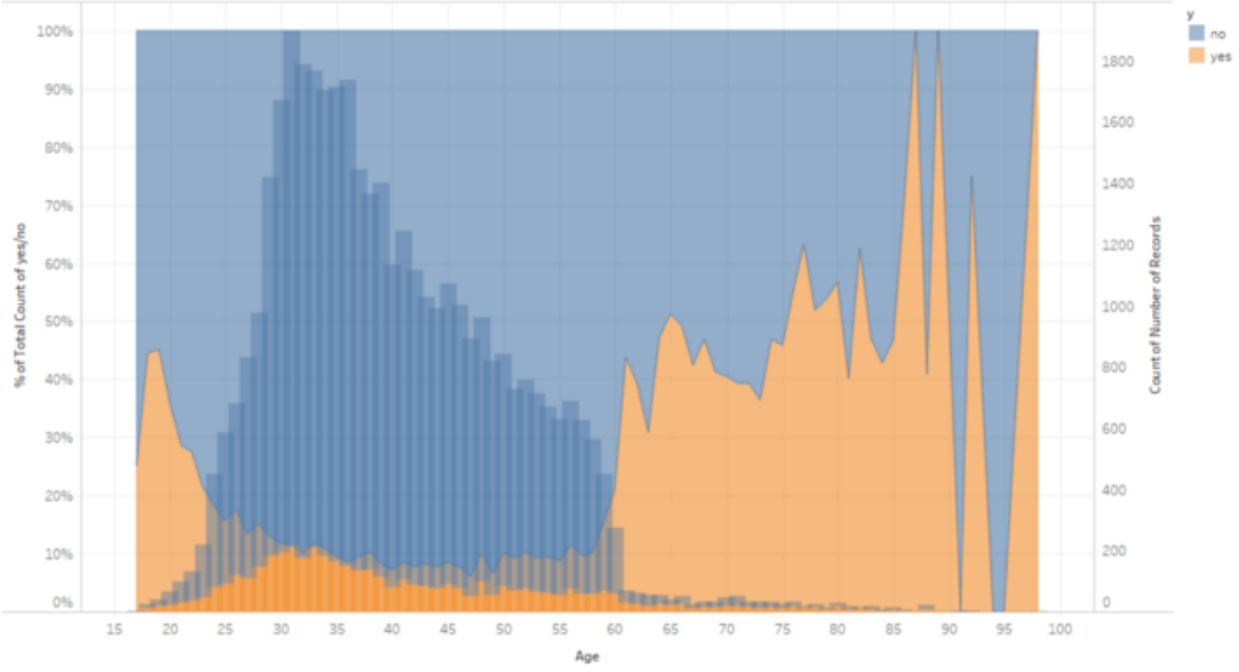
Now, that is what the previous graph looks like with real numbers:

Figure 4. Yes/No Responses by Age



The number of people who responded in the affirmative to the campaign goes up between 23 and 40 on the age scale of Figure 4, but not nearly as dramatically as negative responses do. We saw it in the previous graph that this increase in 'yeses' was completely offset by 'noes' surging in that demographic.

Figure 5. Yes/No Ratio by Age VS Number of Records



In Figure 5, we've joined the last two graphs together. The figure highlights the idea that the area with the lowest yes/no ratio also represents the largest share of records in

the dataset — which must have delivered a double blow to the outcome of the campaign. While 'noes' got support from a very high portion of a really large age group, 'yeses' succeeded with the groups where there simply weren't enough people in the sample. (If you'd like to explore this graph in detail, you should pet this ginger puppy first > 🐕)

Figure 6. Yes/No Ratio by Occupation

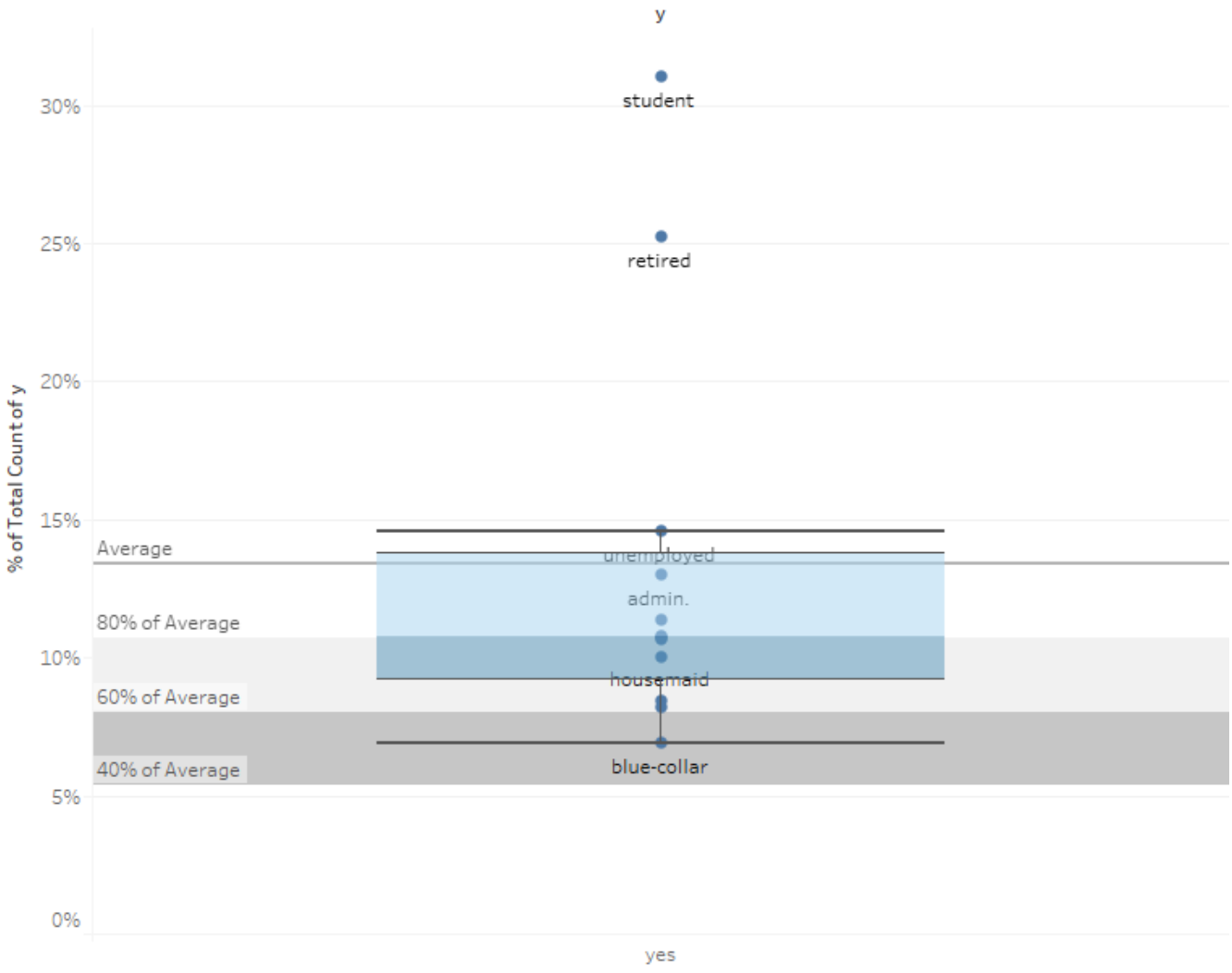
	y	
job	yes	no
student	31,02%	68,98%
retired	25,24%	74,76%
unemployed	14,56%	85,44%
admin.	12,98%	87,02%
management	11,37%	88,63%
unknown	10,76%	89,24%
technician	10,74%	89,26%
self-employed	10,63%	89,37%
housemaid	10,03%	89,97%
entrepreneur	8,41%	91,59%
services	8,18%	91,82%
blue-collar	6,88%	93,12%

The conclusions we drew from Figure 4 find support in this table (Figure 6) as well. As you can see, there are two classes in the occupation list that stand out in terms of their responses to the campaign, i.e. 'students' and 'retired'. Those classes relate to the two peaks in Figure 4: young people and those aged 60 years or over. We've suggested that the latter may have something to do with people starting to retire in big numbers

around that age. The insight we get from this table corroborates that there must be some level of causation indeed.

The box-and-whisker plot with the average line (Figure 7) underscores it once more that those two groups of customers have disproportionally high representation in the 'yes' class. But we aren't going to do a more thorough investigation into this point right now.

Figure 7. Yes Probability by Occupation with Average



Although by no means an exploratory data analysis is a sufficient method to find comprehensive answers to the business problems like this, the bottom line is that the analysis we did has proved the dataset to exhibit some interesting trends we should pay attention to. Please read my post 'Machine Learning Classification with Python for Direct Marketing' to see how this classification case can be solved in a more thorough manner with a predictive machine learning model.

Thanks for your attention!!

Data Science Data Analysis Tableau Classification Machine Learning

About Help Legal