Exploratory Data Analysis

Lesson 1: What Is EDA? - Notes

Quick links

Data Is Ubiquitous

Introducing your instructors

Chris: Hi, and welcome to Exploratory Data Analysis, or EDA. In the last few months, I learned about R, a programming language, and EDA with the help of my friends from Facebook. And in this course, I'll teach you how to use R to conduct Exploratory Data Analysis.

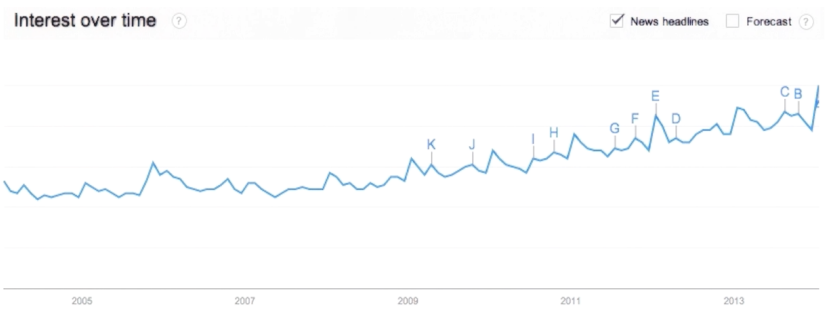
Dean: EDA is one part of the larger process of collecting, learning from and acting on data. In this course, I'll share my advice about working with data and visualizations.

Solomon: By the end of this course, you should feel confident when exploring new data sets to uncover meaningful patterns. In the last lesson, I'll walk you through an analysis of the diamond market with an eye toward building predictive models.

Moira: We're excited to teach this course, and to show you how EDA can be used to answer questions with data. But before we dig in to EDA, let's talk about data. Data is ubiquitous. You can find information about hurricanes, forest fires, and state finances on websites like data.gov. Social networking sites like Facebook, where we work, collect petabytes of data everyday. (One of Facebook's tools, Presto, which is mainly used for adhoc analysis, processes over 1 petabyte of data per day.)And some people have started tracking their own personal data using calendars, mobile apps, and physical activity trackers.

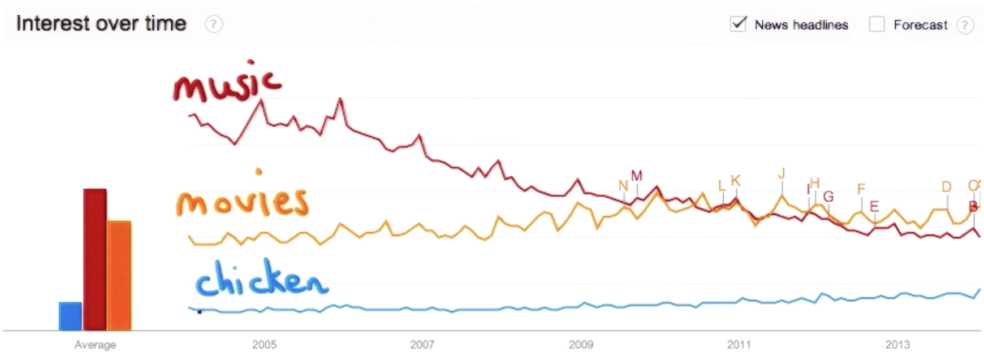
Exploring Google trends

Chris: There is so much data out there in the world, and so little of it has been explored. For example, I was poking around on Google trends and here's what I found. I searched for the word chicken, because I had read a news article that morning about chicken food poisoning and salmonella.



And when I search for the word, Google Trends gave me this graph. This graph shows the interest in searching for the word chicken over time. In reality, it's counting how many times we see chicken in any newspaper headlines in a given month. So from 2005 to 2013, we can see that occurrences of chicken has been increasing. Now, I didn't want to stop my search there, so I decided to enter two more generic word to see what I could find. So I added the word music and I added the word movies. And then I got the graph below.

Quiz: What do you notice in the time series graph below?



Now when you look at this graph, tell me some things that you notice. You can talk about anything interesting that you find in this graph, or you might compare the first graph to this one and write some things that you find. Now, keep in mind there is no right or wrong answer, I just want to get you thinking.

Answer:

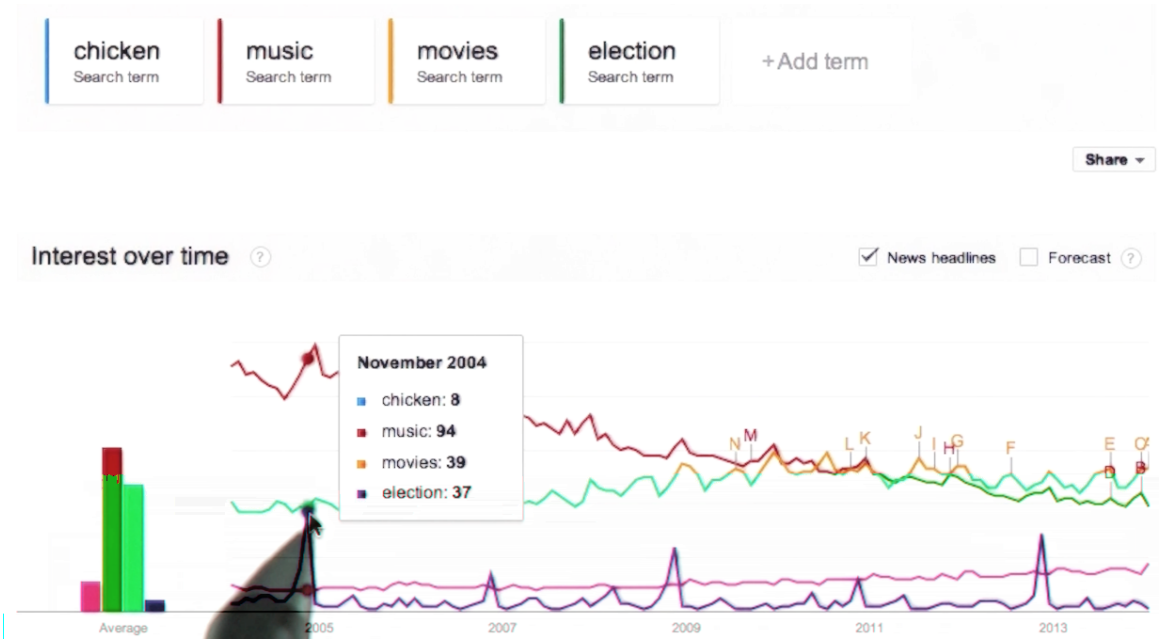
There are a ton of observations to make here, and you might have said something about a particular spike in the graph, or spoken more generally about the trends. Here's what I noticed, I noticed that the vertical scale had changed, since the blue line goes all the way down to the bottom. This means that the number of times that

we see music or movies in headlines is much greater than the number of times we see the word chicken. I also noticed that music used to be more prevalent in headlines, but it's declined in recent years. In fact, the word movies overtook the word music for headline presence around June 2010. That might not have been obvious on the static image. But if you played with the interactive visualization, you could find dates and other interesting things too. If you notice something else that isn't covered in the solution video, share it on the forum.

Go Exploring

Continuing the investigation

I wanted to continue my investigation, so I decided to add another term. This time I used the word election. And here's what was interesting, I noticed these peaks spaced evenly out over time. And then I realized that these peaks made a lot of sense



because the US Presidential election occurs every four years in November. If I hover right over the graph, I can see these peaks in November.

Finally, I wanted to see if there was a word that appeared more than any of these other words in the last few years. So, I typed in the word Facebook. This created a drastic change in the vertical scale in the graph. And I noticed Facebook didn't even come out on the scene until around 2007. I think it's pretty cool that a word such as Facebook dwarfs words like chicken, music and movies. Facebook is certainly the giant when it comes to this graph. This is one small example of how data is

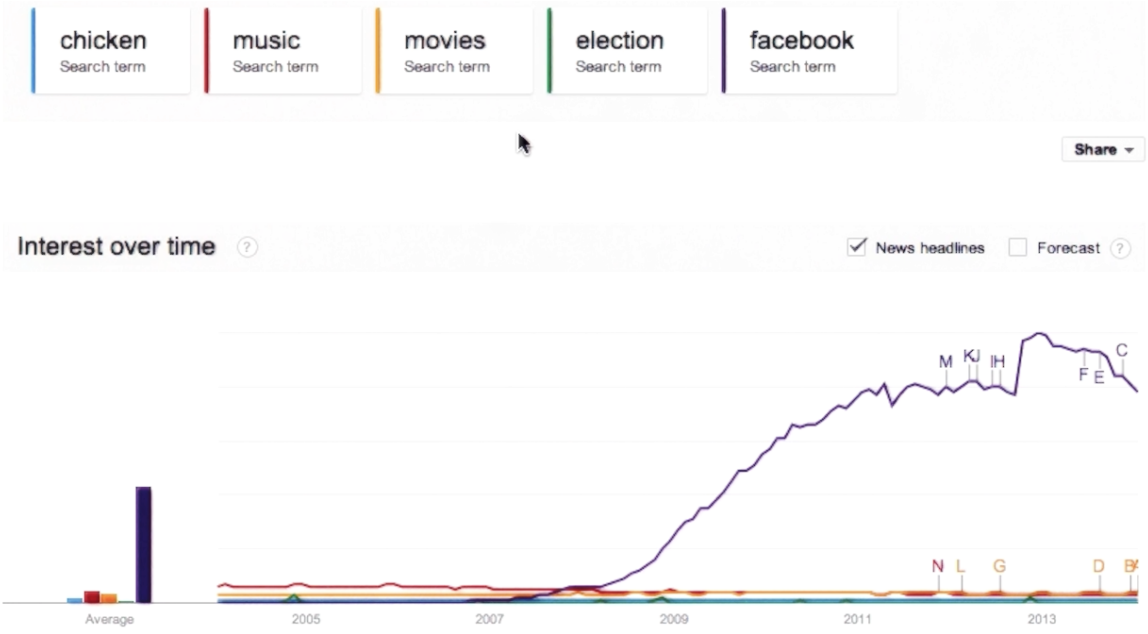
everywhere around us and it's waiting to be explored.

Quiz: Do your own exploration.

Let's have you get your hands dirty, by performing your own trend analysis. Search for celebrities, music, movies, or anything else that you find interesting. You can use the time series graph like I used, or something else on the site. Once you've got your exploration, share it on the forums, and comment on any posts that you find interesting.

So what's getting ubiquitous and cheap?
Data.
And what is complementary to data?
Analysis.

-Hal Varian



Why learn EDA?

Dean: You might be wondering why you should be learning about exploratory data analysis in the first place. Hal Varian, the chief economist at Google, said, “What's

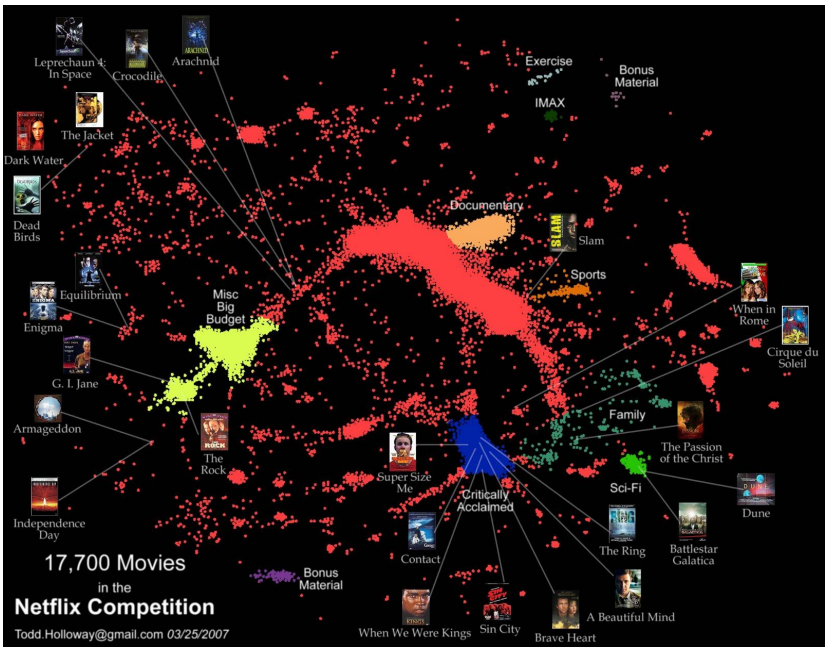
becoming ubiquitous and cheap? Data. And what is complementary to data? Analysis." In general terms, exploratory data analysis or EDA. Is the examination of data and relationships among variables, through both numerical, and graphical methods. It often takes place before more formal, more rigorous statistical analyses.

EDA is often the first part of a larger process. It can lead to insights, or new questions, or even feed into the process of building predictive models. It's also an important line of defense against bad data. It's an opportunity to check some of your assumptions and intuitions about a data set. Many times, business decisions are made using unpolished visualizations that come out of exploratory data analysis. Other times, you might polish some of those visualizations or summaries before presenting them to a larger audience. Let's hear from Chris about an example of how EDA can fit into the larger process of building predictive models.

Netflix Prize Example

Chris: A great example of EDA comes from the Netflix prize competition. Teams competed for a million dollar prize to improve the Netflix movie recommendation system by at least 10%. You can read more about the contest and results by clicking on the following links (Netflix prize dataset visualization, Interactive visualization).

The visualization to the bottom right was created by Todd Holloway to show clusters of movies based on movie ratings. Movies that are closer together, receive similar movie ratings, whether those ratings are high or low. And it's exactly this type of visualization that allows us to get a feel for the data and determine what could be a part of it and what could be explored in more depth.



So, even before performing complex data analyses on the data, we can learn a lot about visualizing our data. Nathan Yau, a noted statistician and visualization guru provides his reflections about this plot. You can read his thoughts and check out the interactive visualization from the instructor notes as well. So, whether you want to learn an in-demand skill or compete in $1,000,000 competitions. EDA has something to offer to you. At the least, learning about EDA will

improve your ability to reason through data, sharpen your communication skills, and expand your career opportunities.

Quiz: Why do YOU want to learn EDA?

Now I want you to take a few minutes to think about your previous work and your educational experiences. Write a few sentences about why you would like to learn about EDA and once you've got your ideas down, share your thoughts and your experiences in the forum. And remember, there's no right or wrong answer here. This question is just for you.

Aude's Interest in Data

Chris: While creating this course, I visited Facebook with Dean, Moira, and Solomon and spoke with many data scientists. I learned a lot about EDA, and I'm excited to share those stories with you throughout the course. First, you'll hear from Aude who will talk about what she finds most interesting about being a data scientist.

Aude: So what I really like is, being able to go from like really raw data - we have like billions and trillions of rows of data. You don't know what's interesting. It's like this gigantic thing and you need to investigate and have some ideas on what might be interesting, and then try investigating it, like, in different ways. Sometimes you use, charts, sometimes you use maps and sometimes you come across something that's interesting and that you want to dig a little bit more into. And then eventually, once we have figured out something interesting that we've developed the corresponding algorithms, then the goal is to have an impact on, either understanding the world or changing the Facebook product. And being able to go from…we have all this data, we don't know what to do with it. So like, we're going to make a significant impact on like, understanding mobility or improving Facebook. That's like a, a real awesome flow of work.

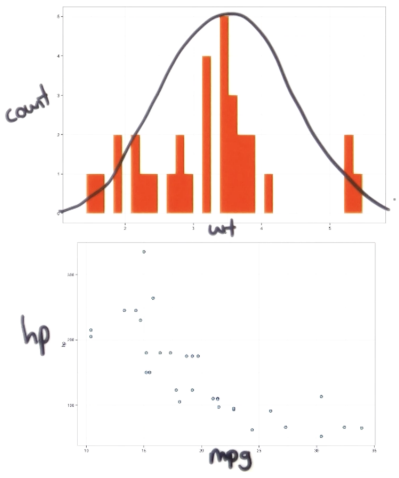
Goals of EDA

So what exactly is exploratory data analysis? You can check out the Wikipedia definition in the instructor notes but simply put it's an approach to understanding data using visualization and statistical tools. Think of EDA as your initial interaction with data.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4

For example, let's say we wanted to look at a set of cars. This data comes from the MT cars data set, which comes with R Studio. Now we couldn't make any sort of conclusions just by reading through the table. It'd be pretty difficult. So instead what we want to do is we want to understand the distribution of the variables for things like miles per gallon, cylinders and perhaps horsepower. We might create histograms, which we'll see later in Lesson three.

The histogram to the right shows the distribution of weights for all the cars. Or we might examine correlations between variables like miles per gallon and horsepower. In that case we'd want to create a scatter plot. We'll examine these in Lesson four.



Our second goal is to assess and validate assumptions. In which future inferences will be based. So we might ask questions like which variables are normally distributed. Or we might be wondering if a variable is biased toward a particular value. Now this might not be the case for our cars data set but certainly in other cases like social behavior or users interacting on the web you might find that.

Third you might want to understand the data before performing and intelligent hypothesis. Well EDA can be the source of an idea for an experiment. It's not a formal

process of hypothesis testing and predictive modeling. Ultimately, we're developing intuition of our data set and how it came into existence. By examining our data we can generate better hypothesis, determine which variables have the most predictive power. And select appropriate Statistical tools, to build our predictive models.

The Growth of Televisions

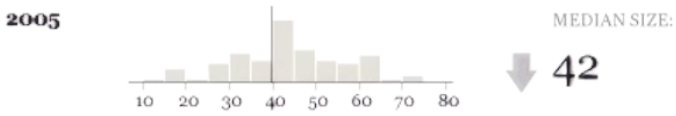
Let's look at another example of EDA. We've linked you to a blog posting on Nathan Yau's blog about Flowing Data. In the post Nathan analyses the size of televisions over the past couple of years. I want you to read through the post and think about how Nathan conducted his EDA. And pay careful attention to the date of visualization and the source of his data. I want you to gather some thoughts about some specific things that Nathan did in his work. And how they relate to the definition and the goals of EDA.

Quiz: Some questions about Nathan's post

According to Nathan's post, what was the median television size in 2006? The largest increase in television size occurred between which two years? What else stands out to you about Nathan's work? What connections are there between his work and the definition and goals of EDA?

Answer:

This video doesn't contain an exhaustive list of everything you might have related to exploratory data analysis. But here are some of the things that we thought about. If you have other ideas, please share your thoughts on the forum. The first thing I notice is that Nathan used histograms to show the distribution of TV sizes in any given year. You can also see that for each year, Nathan included a dark line in the middle. Which represents the median size television for that year. So in 2005, it looks like the median TV size was about 40 inches. So half the TVs were larger and half of them were smaller. Second, Nathan describes where the data originated.

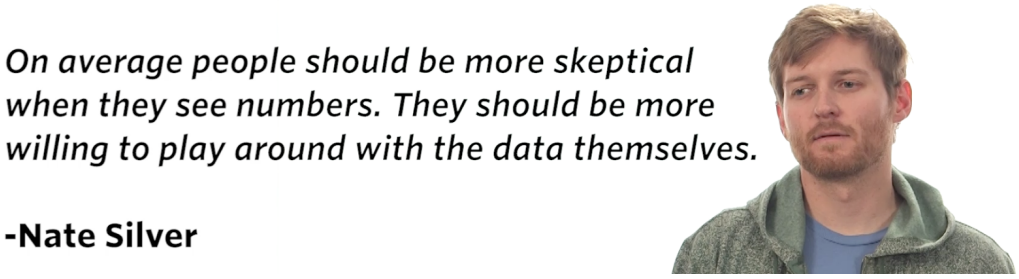


In this case he got his data from CNET Reviews. Nathan describes a caveat in the source of his data. He mentioned that the screen sizes are based on the TVs that were on the market at the time and not on how many people actually bought them. The average television size that people bought could be quite different from the average television size on the market. I think the second part is key and is something we should remember when approaching any data set. We should always be skeptical of what story the data may tell. We should ask questions about our data, inspect it and

consider its context.

Finally, Nathan conducted a time series analysis. So each year after 2002, Nathan indicates whether the median size of televisions increased or whether or not it decreased. In the case of the circle, we know that the television size remained the same. This leaves me wondering what television sizes are doing now. Maybe you can find some data out there and let us know by the end of the course.

Our Approach for This Course



Dean: As you go through this course, and when you're analyzing data on your own, we want you to keep the following in mind. Always be curious and skeptical when you're taking a look at data. Nate Silver a statistician noted for his work in baseball and in politics, captured the sentiment as well. He said, “People should be more skeptical when they see numbers. They should be willing to play around with the data themselves.” Exploratory data analysis is one of the best tools that we have for playing with the data. For letting the data speak really directly. When conducting an exploratory data analysis. We want to test our intuitions about the data set and develop new intuitions.

When you're going through this course and analyzing the data sets that we're going to take a look at. We want you to be excited to play with the data. But we also want you to be open to detecting oddities in the data. Especially through the visualizations and summaries that you produce.

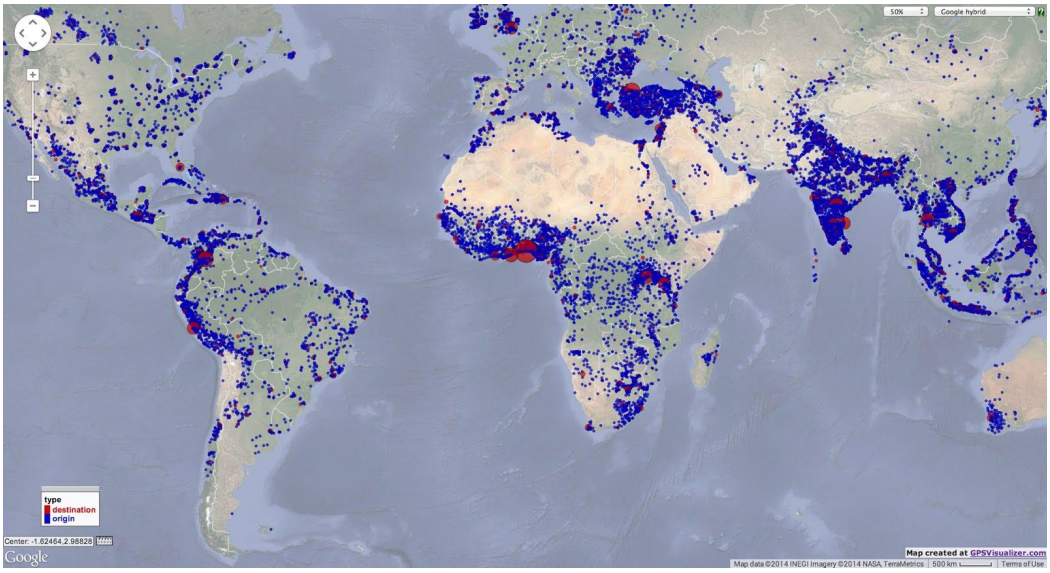
Aude Explores Coordinated Migration

Chris: We want you to develop a mindset of being both curious and skeptical, when you work with data. To help you get into this mindset, I want to share another conversation that I had with Aude. In this next video, I want you to listen to Aude's work and look out for how she demonstrated this

exact mindset.

Aude: So we gathered all the hometowns and current cities from the users and I was looking at conditional probabilities given a hometown. What is the probability that you currently live in each different cities? Like, for example given that your hometown is New York, what is the probability that you live in Chicago or that you still live in New York or that you live in San Francisco or Paris and so on. And what I was expecting is that, at least, the most likely city, where you would live right now would be your hometown.

If you grew up in Chicago, the most likely place that you're going to be now is still Chicago. You could be moving as well but the most likely place would remain your hometown. But I saw a fair number of cases where the most likely current city was different from the hometown and that was, was a fairly high probability. I was really

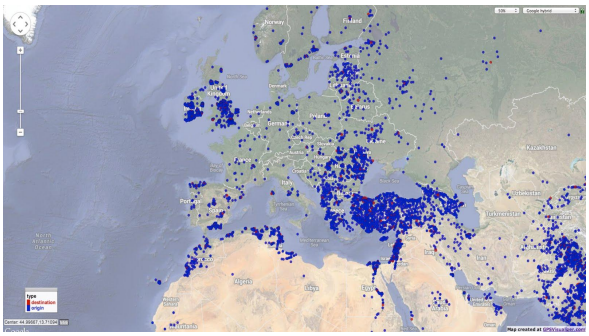


surprised. I was wondering if I had a problem in my computations, if there was some issues upstream of what I was doing.

So I decided to put all the cities on a map. All the pairs of hometowns and current cities for which the most likely current city was different from the hometown. And what we saw on this map was really fascinating because it was really not what we we're expecting. It was not a bug in the code. We were really seeing patterns arise.

Here we only plotted pairs of hometown and current city, so there's no movement between the pairs but what we see is that a lot of these cities for which the most likely current city is different from the hometown arise in western Africa or in India or in like

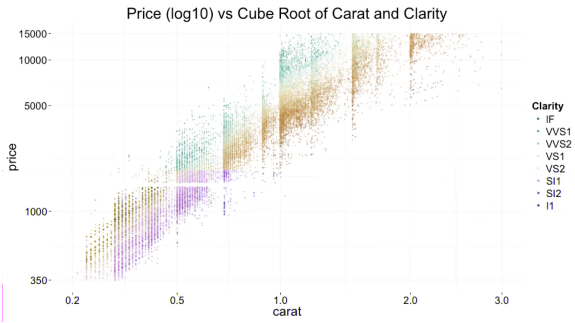
Turkey, which we were not necessarily expecting at the beginning. And there were a lot of small cities all moving to the same current city and so we decided to dig a bit more into it.



One thing that happens is that sometimes the distribution of the current city is very flat. Given that you grew up in, let's say Paris, maybe you're still living in Paris but maybe you live in like one of the thousand cities in the suburbs and so the distribution is really flat and so we have to decide what was considered as a coordinates demarcation. We decided yeah, the probability to move to that city is high enough that we're considering that.

And the other thing we have to think about is that if you look at the map at the world scale or if you zoom to a very specific area, you don't want to see the same things. So,we also want it to have interactivity in the visualization. And so we decided to use D3, which is a Javascript-based visualization framework, which enables you to have a lot of interactivity with with your data and enabled us to do a lot of that exploration and so on. (You learn more about Aude's work here or about global urbanization trends here).

Course Overview



Chris: Visualizations like this one are a core component of exploratory data analysis. We're going to be using the R programming language to create these types of visualizations and to explore data throughout the course. You can read more about the R programming language and how it was developed by clicking on the link. We'll also learn about RStudio. It's a graphical user interface for programming in R. You'll learn all about RStudio in the next lesson, and future lessons will cover visualization and statistical techniques for data exploration. We'll start by examining one variable at a time in a dataset. This will give us an idea of its underlying distribution and what type of values the variable takes on. Eventually our visualizations will become more complex as we look at pairs of variables and multiple variables all at once. At the end of the course you'll have the chance to

conduct your own exploratory data analysis. I hope you're getting excited to explore data. See you in the next lesson.