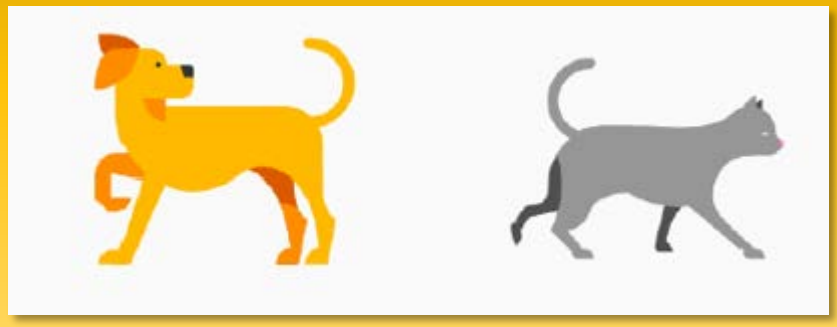
DATA SCIENCE BASIC



Data Science Basic

In this session

- The 5 questions data science answers
- Is your data ready for data science
- Ask a question you can answer with data
- Predict an answer with a simple model

Data Science Basic

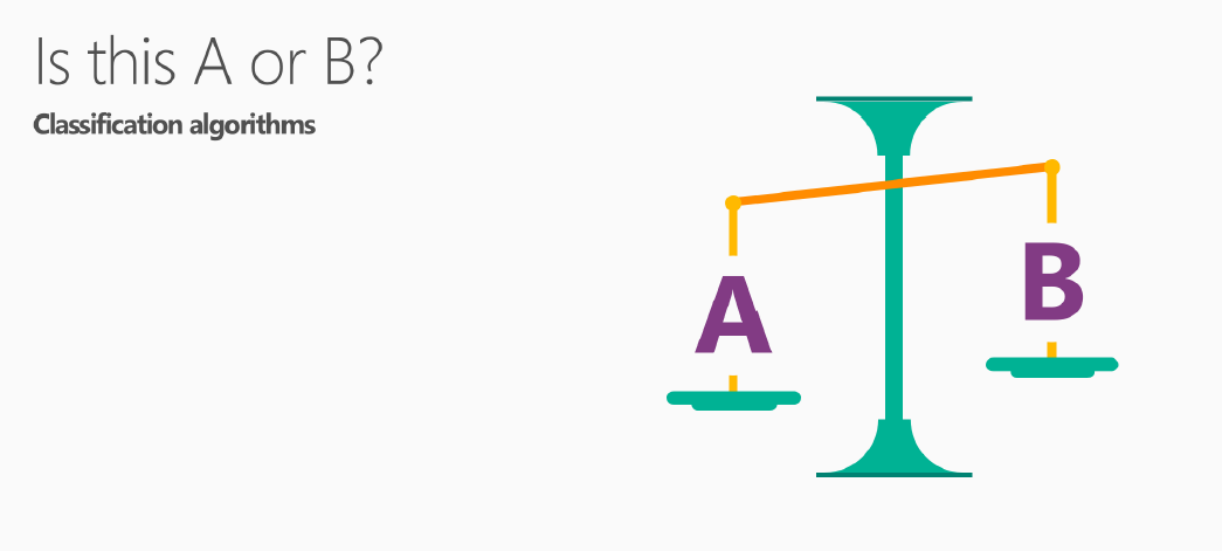
The 5 questions data science answers

The 5 questions data science answers

- Is this A or B?
- Is this weird?
- How much – or – How many?
- How is this organized?
- What should I do next?

Data Science Basic

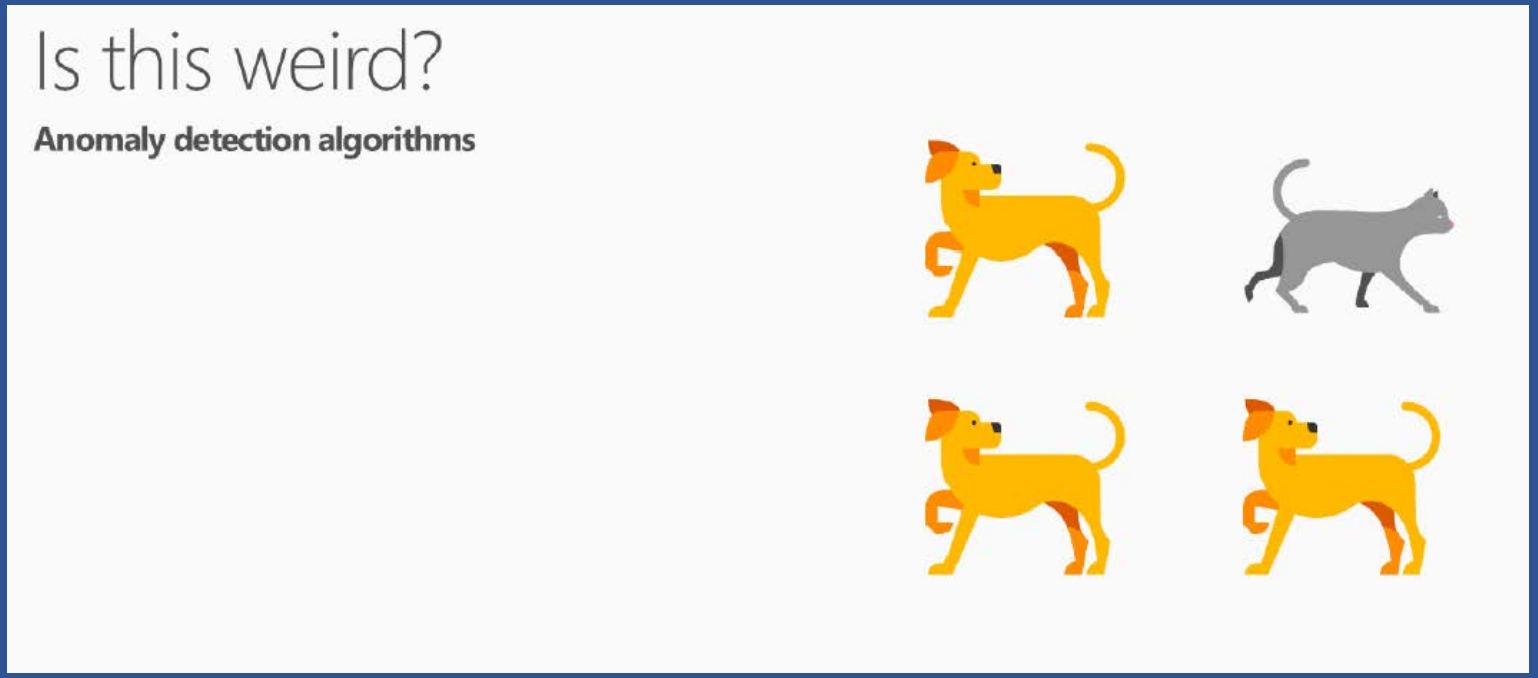
The 5 questions data science answers



- Will this tire fail in the next 1,000 miles: Yes or no?
- Which brings in more customers: a $5 coupon or a 25% discount?
- Can also be more than two options: Is this A or B or C or D, etc.?
- Classification algorithms: helps choosing the most likely one.

Data Science Basic

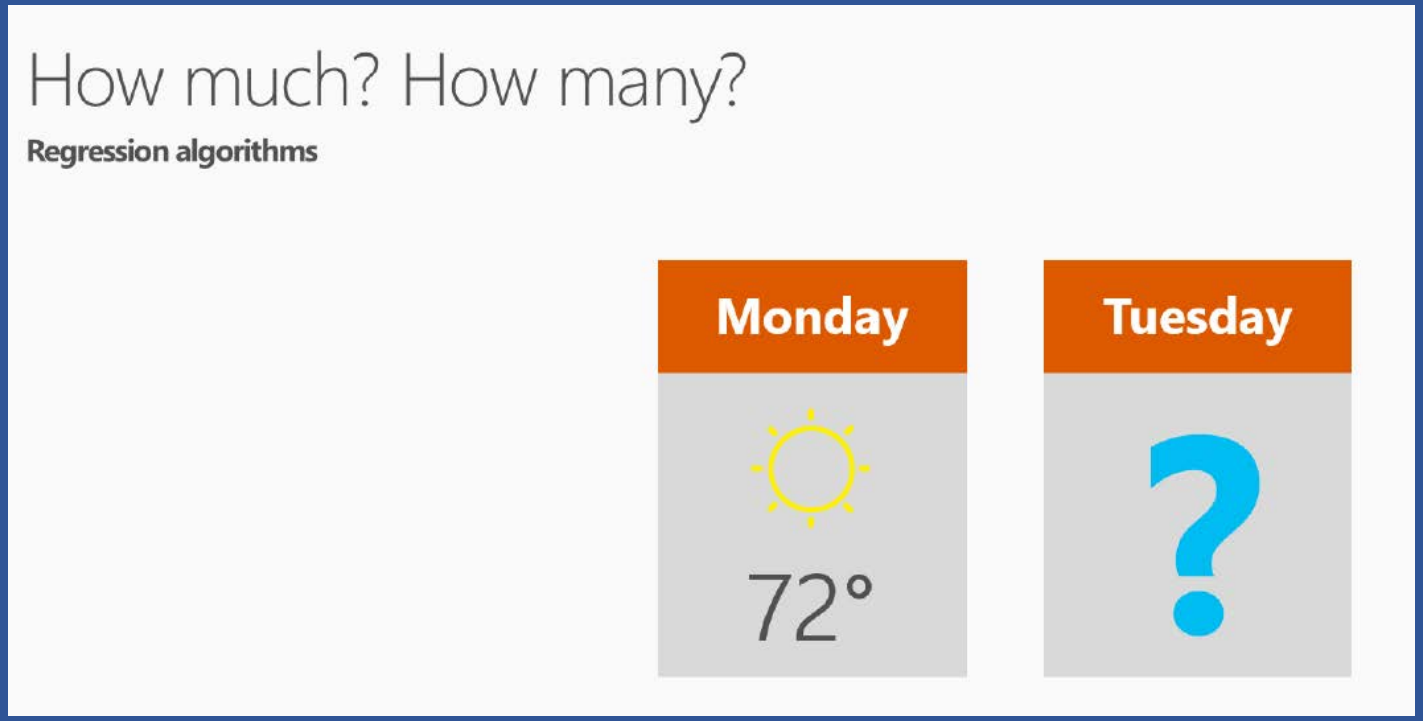
The 5 questions data science answers



- Is this pressure gauge reading normal?
- Is this message from the internet typical?
- Credit card purchase pattern normal?
- Anomaly algorithms: Detect unexpected or unusual events or behaviors

Data Science Basic

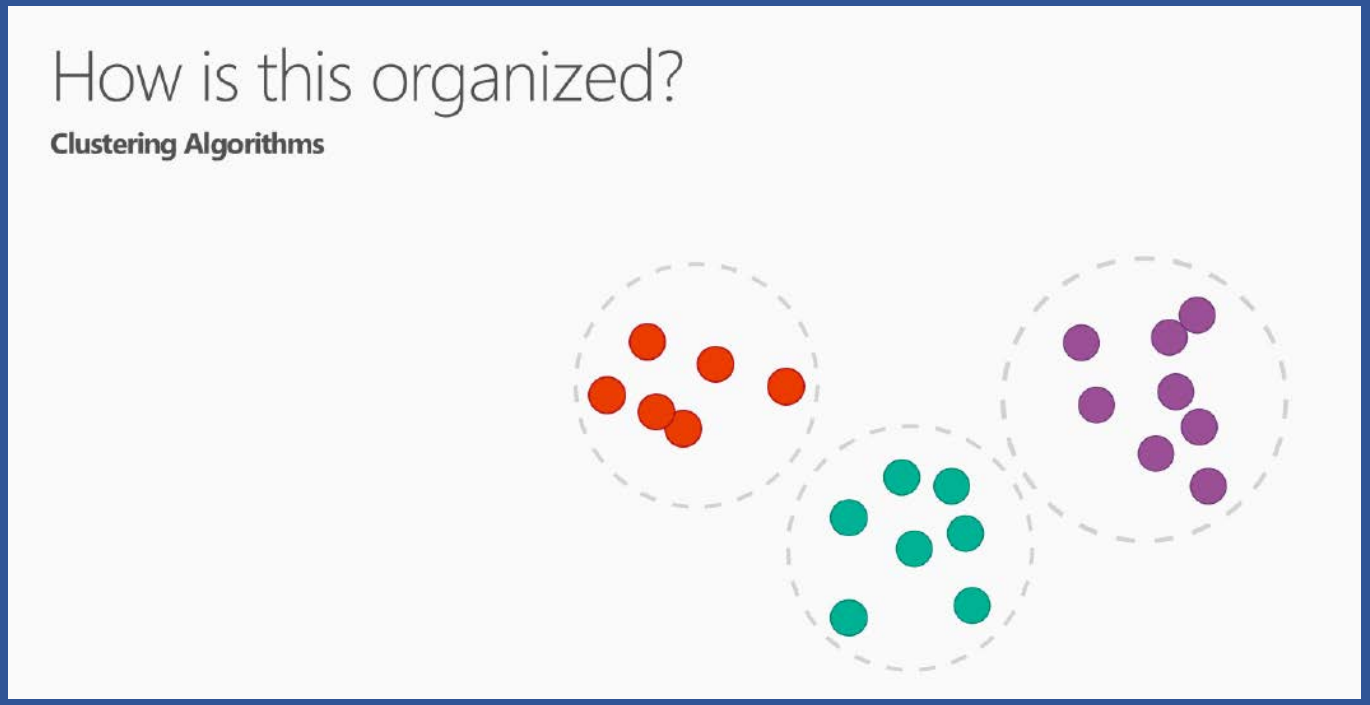
The 5 questions data science answers



- What will the temperature be next Tuesday?
- What will my fourth quarter sales be?
- Regression algorithms: Good for question that asks for a number

Data Science Basic

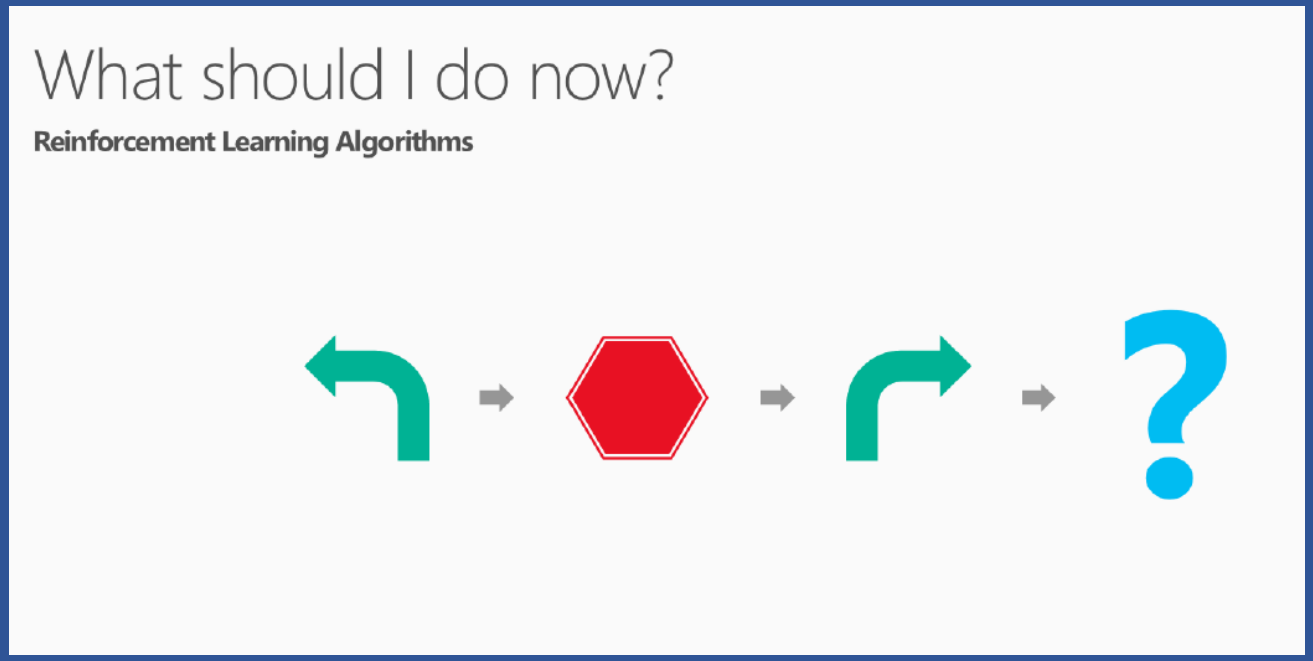
The 5 questions data science answers



- Which viewers like the same types of movies?
- Which printer models fail the same way?
- Clustering algorithms: helps arranging data into groups
- Understanding how data is organized, helps predict behaviors and events.

Data Science Basic

The 5 questions data science answers



- Adjust the temperature or leave it where it is?
- At a yellow light, brake or accelerate?
- Keep vacuuming, or go back to the charging station?
- Reinforcement learning algorithms: the brains of rats and humans respond to punishment and rewards. learning from trial and error.

Data Science Basic

Is your data ready for data science

We need data that is:

- Relevant
- Connected
- Accurate
- Enough to work with

Data Science Basic

Is your data ready for data science

Relevant

Irrelevant Data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant Data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

- We need to know Blood alcohol content %
- Price of milk and Red Sox are irrelevant

Data Science Basic

Is your data ready for data science

Connected

Disconnected Data

Grill temp. (Fahrenheit)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

Connected Data

Grill temp. (Fahrenheit)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

- Quality of hamburgers
- But notice the gaps in the table on the left
- It's common to have holes like this

Data Science Basic

Is your data ready for data science

Accurate

- Top left: precise=yes/accurate=no
- Button left: precise=no/accurate=no
- Top right: precise=yes/accurate=yes
- Button right: precise=no/accurate=yes

Data Science Basic

Is your data ready for data science

Enough to work with



We need enough data to work with

1. Not enough data: can not make decision
2. Barely enough data: can make basic decision (Is it somewhere I might want to visit? It looks bright, that looks like clean water – yes, that's where I'm going on vacation.)
3. Enough data: can make detailed decision (Now I can look at the three hotels on the left bank. You know, I really like the architectural features of the one in the foreground. I'll stay there, on the third floor.)

Data Science Basic

Ask a question you can answer with data

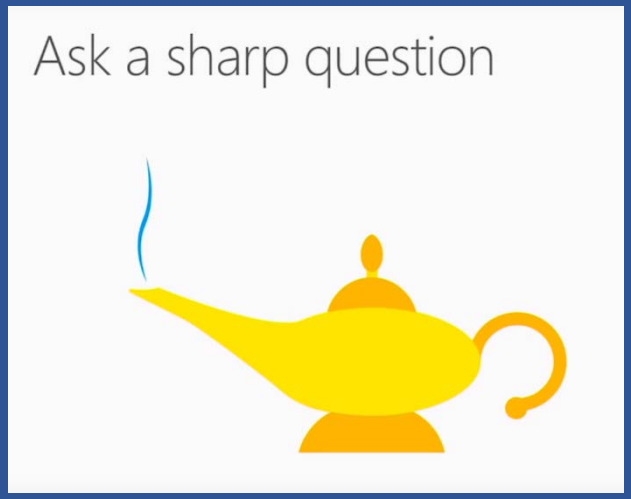
Ask a question you can answer with data

- Sharp question is the Key
- Know Target Data
- Reword the question

Data Science Basic

Ask a question you can answer with data

Sharp question is the Key



- Asking a sharp question is the most important
- ML is a mischievous genie
- "What's going to happen with my stock?", the genie might answer, "The price will change"
- "What will my stock's sale price be next week?", the genie can't help but give you a specific answer and predict a sale price

Data Science Basic

Ask a question you can answer with data

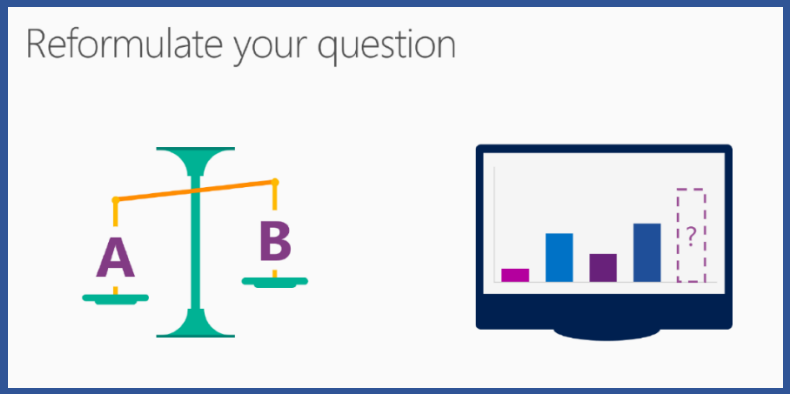
Know Target Data

- Target data = what we are trying to predict
- Must have examples of the target in data.
- Question = "What will my stock's sale price be next week?" Target = stock price history.
- Question = "Which car in my fleet is going to fail first?" Target = previous failures data.

Data Science Basic

Ask a question you can answer with data

Reword the question



- Question dictates the algorithm
- "Is this data point A or B?" = classification
- "How much?" or "How many?" = regression
- "Which news story is the most interesting to this reader?"
- Algorithm = classification A or B or C or D; difficult
- Reword = "How interesting is each story on this list to this reader?"
- Give each article a numerical score
- Identify the highest-scoring article; easy
- Above example change classification question into a regression question

Data Science Basic

Predict an answer with a simple model

Predict an answer with a simple model

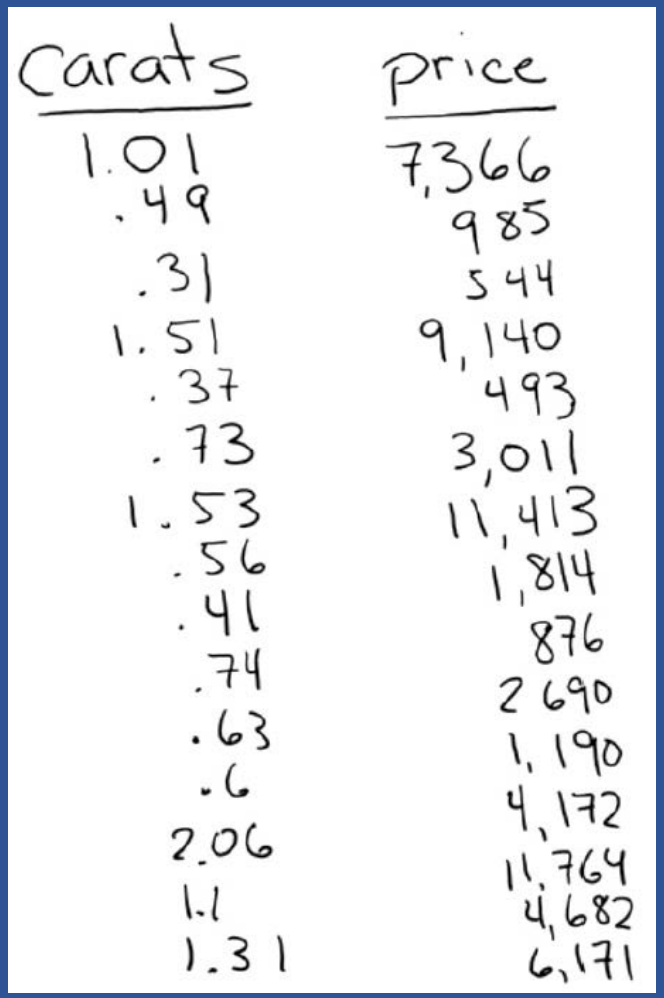
- Collect relevant, accurate, connected, enough data
- Ask a sharp question
- Plot the existing data
- Draw the model through the data points
- Use the model to find the answer
- Create a confidence interval

Data Science Basic

Predict an answer with a simple model

Collect relevant, accurate, connected, enough data

- I want to how much 1.35 carat diamond will cost
- Go to jewelry store
- Write down the price of all of the diamonds
- List has two columns
- Each column has a different attribute
- Weight in carats and price
- Each row is a single data point
- Data that represents a single diamond.
- This is a small data set; a table



Carats	price
1.01	7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2690
.63	1,190
.6	4,172
2.06	11,764
1.1	4,682
1.31	6,171

Data Science Basic

Predict an answer with a simple model

This data set meets our criteria for quality:

- Relevant: weight is definitely related to price
- Accurate: we double-checked the prices that we write down
- Connected: there are no blank spaces in either of these columns
- Enough data: to answer our question

Data Science Basic

Predict an answer with a simple model

Ask a sharp question

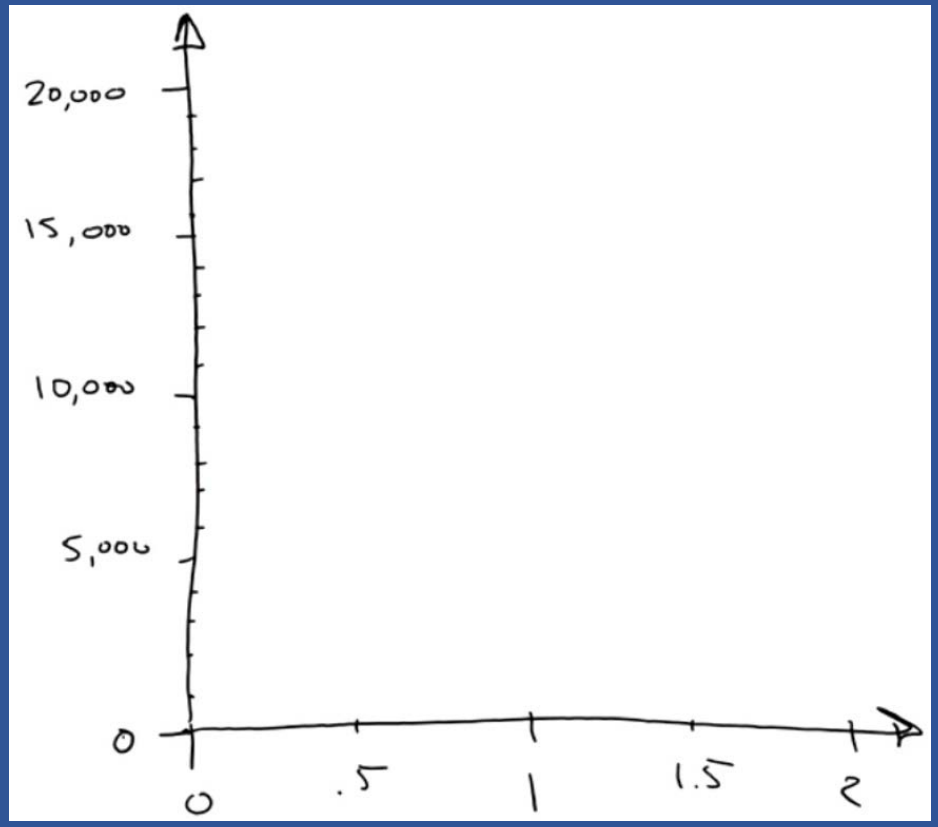
- How much will it cost to buy a 1.35 carat diamond?
- Our list doesn't have a 1.35 carat diamond
- Use the rest of our data to get an answer to the question

Data Science Basic

Predict an answer with a simple model

Draw axis

- Draw a horizontal number line, called an axis, to chart the weights
- The range of the weights is 0 to 2
- Line covers that range and put ticks for each half carat
- Draw a vertical axis to record the price and connect it to the horizontal weight axis
- This will be in units of dollars
- Now we have a set of coordinate axes.

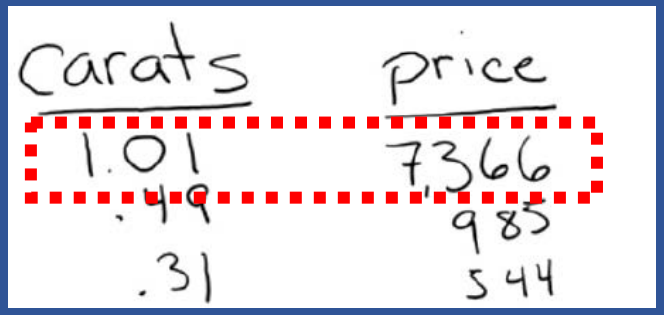


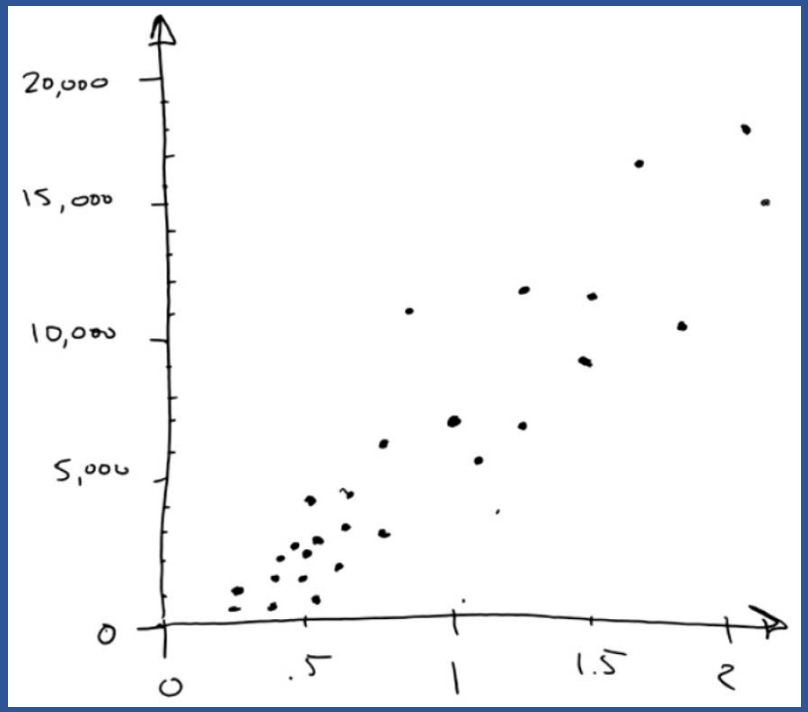
Data Science Basic

Predict an answer with a simple model

Plot the existing data

- Make a scatter plot
- Great way to visualize numerical data sets
- For the first data point, we eyeball a vertical line at 1.01 carats. Then, we eyeball a horizontal line at $7,366. Where they meet, we draw a dot
- This represents our first diamond.
- Now we go through each diamond on this list and do the same thing.
- We get a bunch of dots, one for each diamond



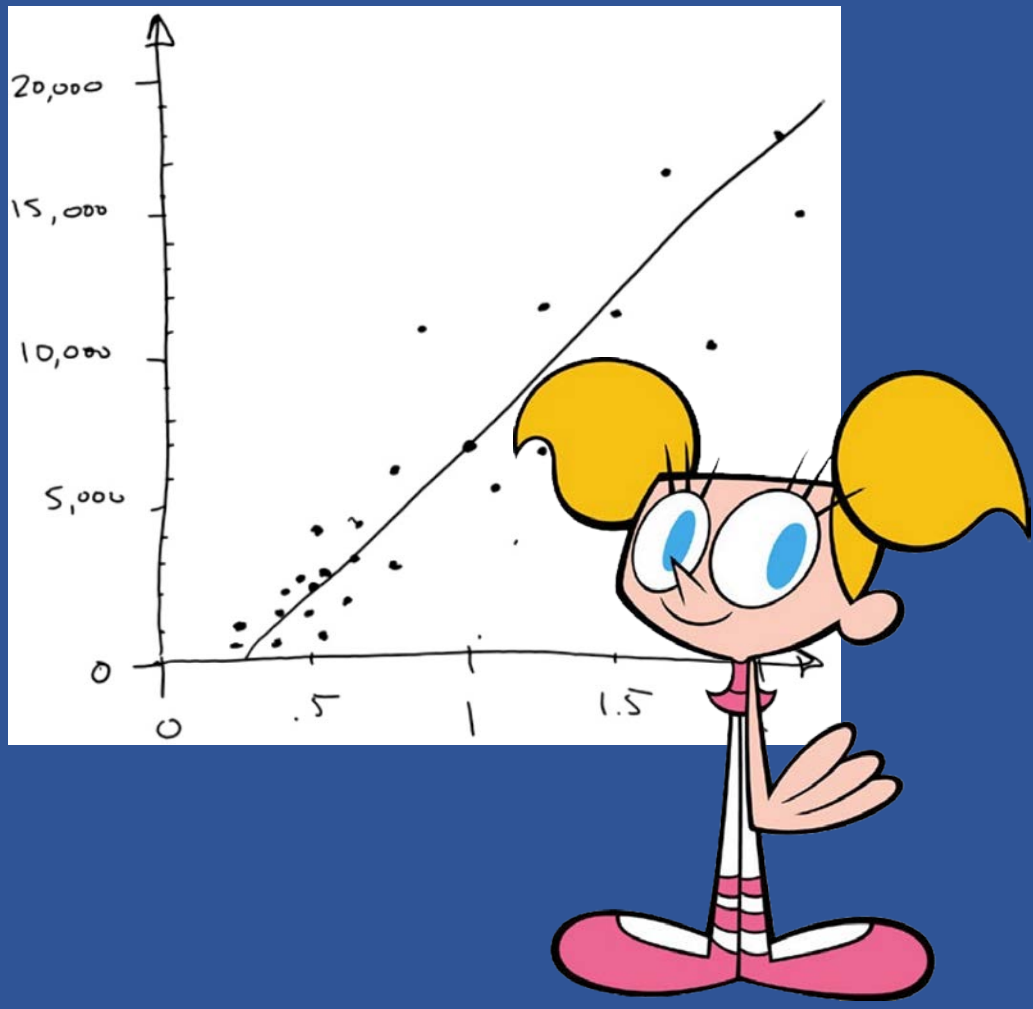


Data Science Basic

Predict an answer with a simple model

Draw the model through the data points

- Look at the dots and squint, the collection looks like a fat, fuzzy line
- Draw a straight line through it
- This a model
- Model = cartoon
- The cartoon is wrong
- But, it’s a useful simplification
- The line doesn't go through all the data points.
- It has some noise or variance
- But, it's a useful simplification
- Question = How much? regression
- we're using a straight line, linear regression

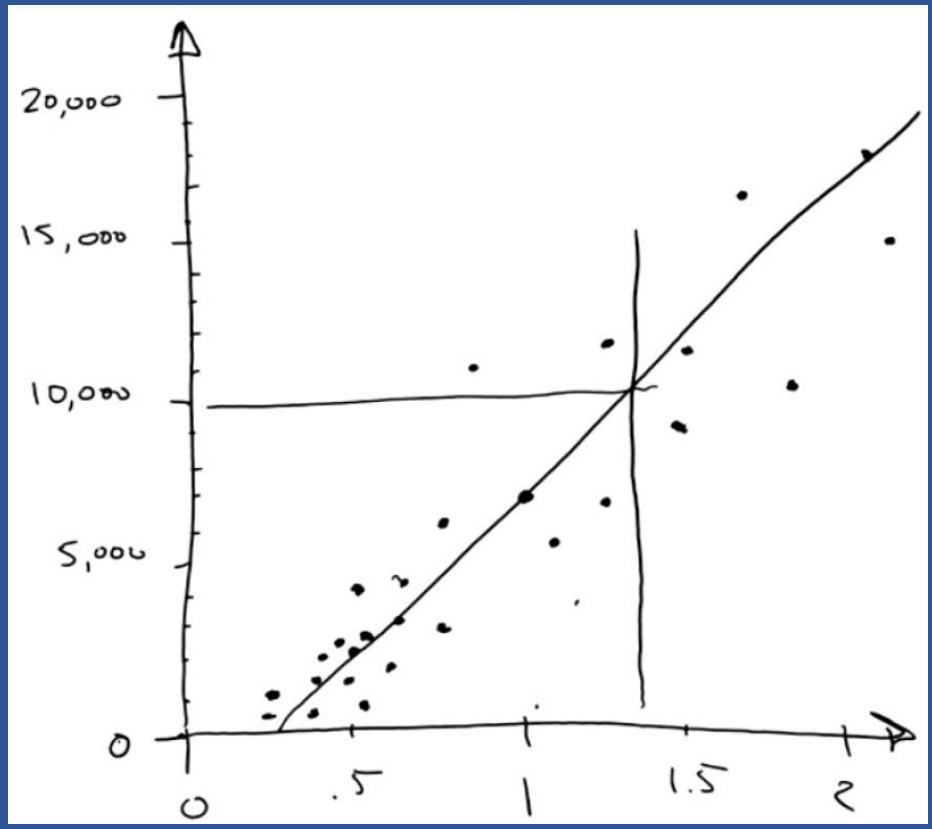


Data Science Basic

Predict an answer with a simple model

Use the model to find the answer

- How much will a 1.35 carat diamond cost?
- Look at 1.35 carats
- Draw a vertical line
- Draw at horizontal line to the dollar axis
- It hits right at 10,000
- Answer = about $10,000

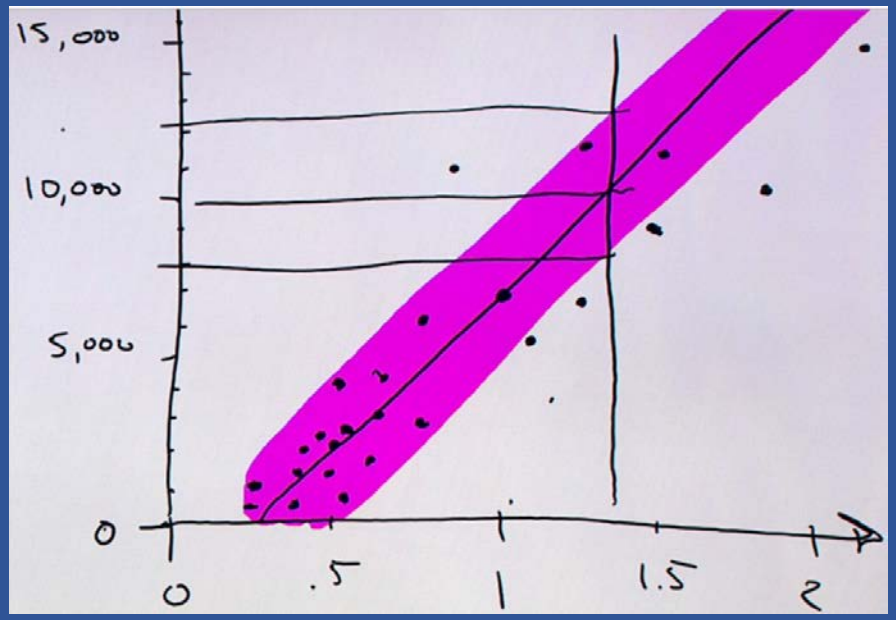


Data Science Basic

Predict an answer with a simple model

Create a confidence interval

- How precise this prediction is?
- Is it a lot higher or lower?
- Draw an envelope around the regression line
- that includes most of the dots.
- This envelope is called our confidence interval
- We're pretty confident that prices fall within this envelope, because in the past most of them have.
- We can draw two more horizontal lines from where the 1.35 carat line crosses the top and the bottom of that envelope.
- The price of a 1.35 carat diamond is about $10,000 - but it might be as low as $8,000 and it might be as high as $12,000



Data Science Basic

Predict an answer with a simple model

We're done, with no math or computers!!

- We did what data scientists get paid to do, and we did it just by drawing:
- We asked a question that we could answer with data
- We built a model using linear regression
- We made a prediction, complete with a confidence interval

And we didn't use math or computers to do it.

Data Science Basic

Predict an answer with a simple model

Now if we'd had more information, like...

- the cut of the diamond
- color variations (how close the diamond is to being white)
- the number of inclusions in the diamond

...then we would have had more columns. In that case, math becomes helpful. If you have more than two columns, it's hard to draw dots on paper. The math lets you fit that line or that plane to your data very nicely.

Also, if instead of just a handful of diamonds, we had two thousand or two million, then you can do that work much faster with a computer.

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

Data Science Basic

More information

More information on Data Science Basic

An introduction to Data Science: Jeffrey Stanton
https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf