Why data scientists need to know how to KISS*

Alon Nir
PyData Dublin Meetup
13/7/20

* Keep It Simple Sweetie

Agenda

1. Introduce the German Tank Problem
2. Apply the technique to a modern-day setting (shared cars in Tel-Aviv)
3. Tell you why you should care about German tanks and cars in TLV

About Me

- Senior data scientist and a data science lead
- From Tel-Aviv, now in London
- Sometimes I share interesting things on linkedin or twitter: @alonnir

1. The German Tank Problem

German Tank Problem

In WWII, the Allies wanted to know how many tanks the Germans were producing.

Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

In WWII, the Allies wanted to know how many tanks the Germans were producing.

Month	Intelligence estimate
June 1940	1,000
June 1941	1,550
August 1942	1,550

Source: German Tank Problem on Wikipedia
Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

In WWII, the Allies wanted to know how many tanks the Germans were producing.

Month	Intelligence estimate	Statistical estimate
June 1940	1,000	169
June 1941	1,550	244
August 1942	1,550	327

Source: German Tank Problem on Wikipedia
Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

In WWII, the Allies wanted to know how many tanks the Germans were producing.

Month	Intelligence estimate	Statistical estimate
June 1940	1,000	169
June 1941	1,550	244
August 1942	1,550	327

??

Source: German Tank Problem on Wikipedia
Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

In WWII, the Allies wanted to know how many tanks the Germans were producing.

Month	Intelligence estimate	Statistical estimate	German records
June 1940	1,000	169	122
June 1941	1,550	244	271
August 1942	1,550	327	342

Source: German Tank Problem on Wikipedia
Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

So, how did the statisticians do it?

Parts on captured tanks indicated that each tank has a serial number, and it's just a running, sequential number (1,2,3,4....).

Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

German Tank Problem

So, how did the statisticians do it?

Parts on captured tanks indicated that each tank has a serial number, and it’s just a running, sequential number (1,2,3,4....).

In a frequentist approach, the MVUE (minimum-variance unbiased estimator) is pretty straightforward..

Image credit: adapted from image by Bundesarchiv, Bild 183-H26258 / CC-BY-SA 3.0, CC BY-SA 3.0 de, on Wikipedia

Statistical Approach

$$N = m + m/k - 1$$

where:

N = Estimated number of tanks
m = largest number observed
k = number of items observed

Statistical Approach

$$\hat{N} = \underbrace{m}_{\substack{\text{largest}\\ \text{number}\\ \text{observed}}} + \underbrace{m/k - 1}_{\substack{\text{Average gap}\\ \text{between}\\ \text{observations}}}$$

N^ = Estimated number of tanks
m = largest number observed
k = number of items observed

2. Shared Cars in Tel-Aviv

Shared Cars in Tel-Aviv

In 2017 the city of Tel-Aviv launched a shared car programme, called AutoTel.

One interesting thing about the cars is...

Image credit: adapted from image by David Shai on Wikipedia, CC BY-SA 4.0

Shared Cars in Tel-Aviv

Image credit: adapted from image by David Shai on Wikipedia, CC BY-SA 4.0

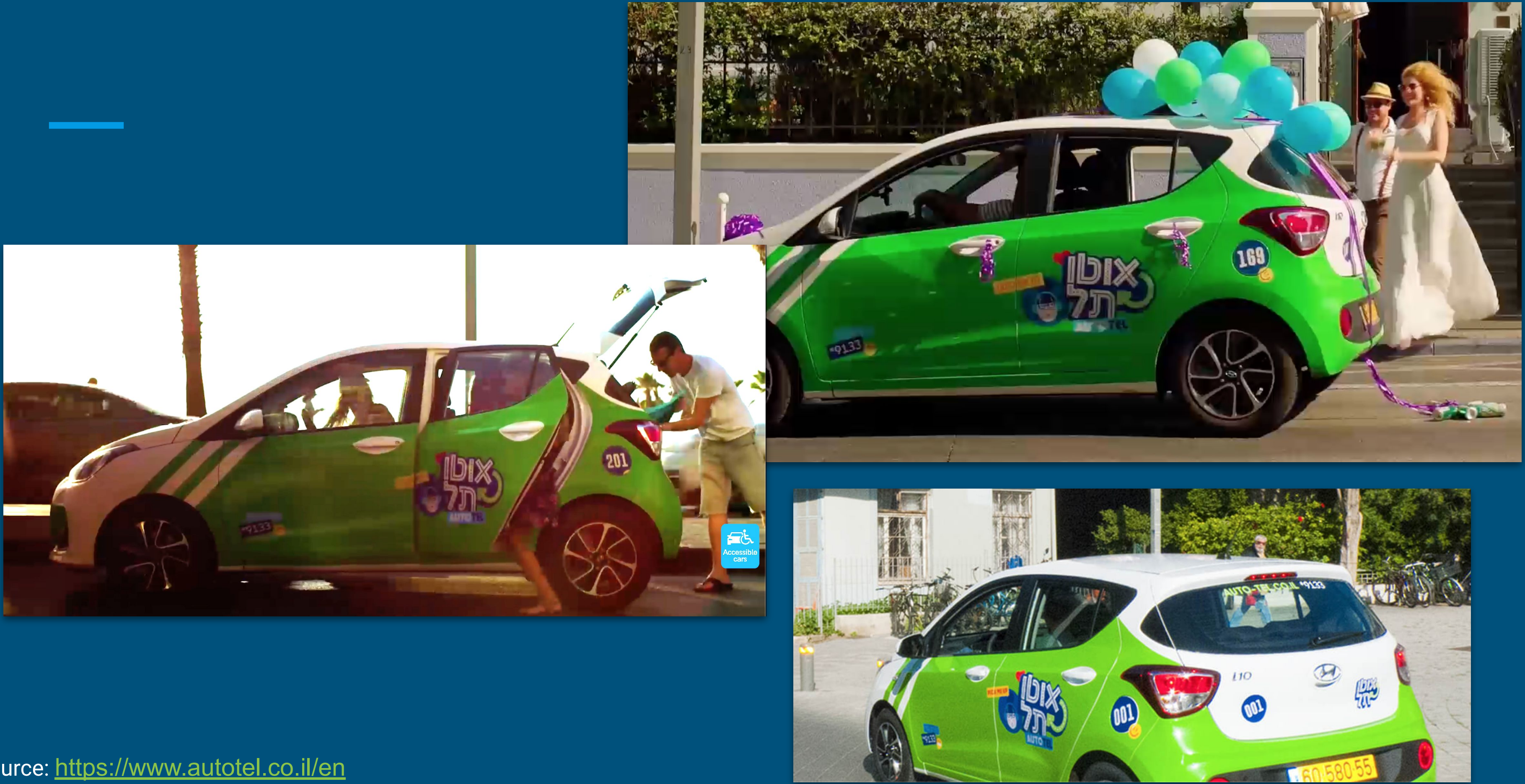
Shared Cars in Tel-Aviv

Can we discover how many shared cars run in Tel-Aviv using those stickers?

Shared Cars in Tel-Aviv

Can we discover how many shared cars run in Tel-Aviv using those stickers?

Yes. And we don’t have to write a single line of code.



Source: https://www.autotel.co.il/en

Shared Cars in Tel-Aviv

Cars observed: 1, 169, 201

$m = 201$

$k = 3$

$$\rightarrow \hat{N} = 201 + 201/3 - 1$$
$$= 201 + 67 - 1$$
$$= 267$$

Shared Cars in Tel-Aviv

Cars observed: 1, 169, 201

m = 201

k = 3

→ $\hat{N}$ = 201 + 201/3 - 1
= 201 + 67 -1
= 267

Actual number: **273** (~2.2% off)

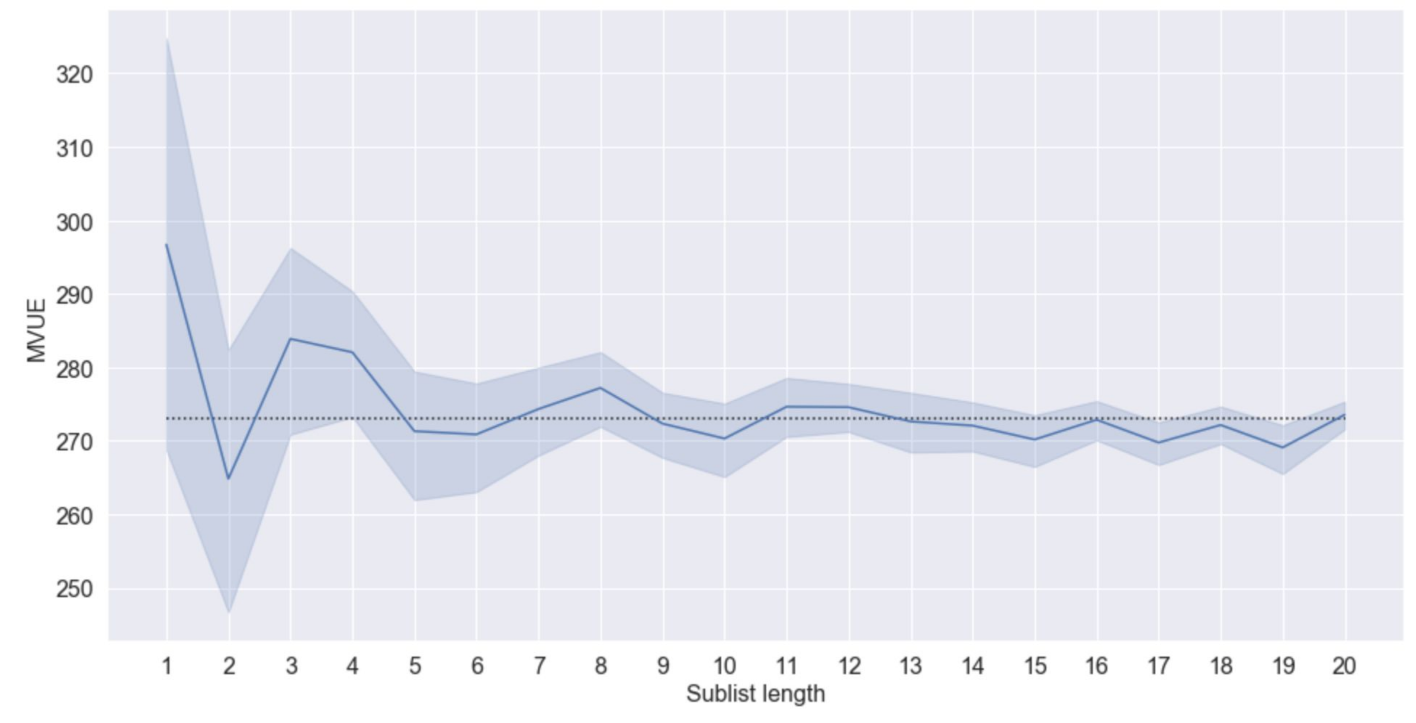
Shared Cars in Tel-Aviv

At this point you might be wondering:

- Maybe it’s just a fluke?
- This is PyData. Where’s the Python?!

(switch to notebook)

Shared Cars in Tel-Aviv



Shared Cars in Tel-Aviv

Follow ups:

- How would this scale if we had 2,730 cars, 27,300 cars or 273,000 cars?
- Bayesian approaches

3. Takeaway / Why should you care?

Takeaway

We’ve seen a simple, ~80 y/o technique did well on a modern problem.

Takeaway

We’ve seen a simple, ~80 y/o technique did well on a modern problem.

- Don’t fall in love with the tool /
  get distracted by the shiny object syndrome
  → **KISS**

- Focus on the research/business question

Thank You!

/alonnir on linkedin, twitter and github