Announcements

- Readings (Week 5 due Wednesday 2/3)
  - In Slack!

- Assignment 2:
  - By end of week 5 (Friday 2/5)
  - Text processing
  - It is OPEN ENDED!!!

Questions for Data Science

Data Science: the easy way

- Dataset
  - Structured, well formatted
  - Minor issues in the data (like in SDPD)
- Question
  - You have precise guidelines on what to look for
- Your job is
  - code for the analysis
  - Present the results

vehicle_stops_2016_datasd

stop_id	stop_cause	service_area	subject_race	subject_sex	subject_age	timestamp	stop_date	stop_time	sd_resident	arrested	searched
1308198	Equipment Violation	530	W	M	28	2016-01-01 00:06:00	2016-01-01	0:06	Y	N	N
1308172	Moving Violation	520	B	M	25	2016-01-01 00:10:00	2016-01-01	0:10	N	N	N
1308171	Moving Violation	110	H	F	31	2016-01-01 00:14:00	2016-01-01	0:14			
1308170	Moving Violation	Unknown	W	F	29	2016-01-01 00:16:00	2016-01-01	0:16	N	N	N
1308197	Moving Violation	230	W	M	52	2016-01-01 00:30:00	2016-01-01	0:30	N	N	N
1308200	Moving Violation	710	H	M	24	2016-01-01 00:30:00	2016-01-01	0:30	Y	N	N
1308174	Moving Violation	Unknown	O	M	20	2016-01-01 00:35:00	2016-01-01	0:35	Y	N	N
1308199	Moving Violation	440	H	M	50	2016-01-01 00:45:00	2016-01-01	0:45	Y	N	N
1308979	Moving Violation	310	H	F	25	2016-01-01 01:03:00	2016-01-01	1:03	Y	N	Y
1308965	Moving Violation	240	W	F	23	2016-01-01 01:10:00	2016-01-01	1:10	Y	N	N
1308175	Moving Violation	120	O	M	54	2016-01-01 01:20:00	2016-01-01	1:20	Y	N	N
1308176	Moving Violation	520	W	F	53	2016-01-01 01:39:00	2016-01-01	1:39	Y	N	N
1308177	Moving Violation	520	W	M	35	2016-01-01 01:57:00	2016-01-01	1:57	N	N	N
1308178	Moving Violation	520	W	M	29	2016-01-01 02:00:00	2016-01-01	2:00	N	Y	N
1308180	Moving Violation	510	B	M	38	2016-01-01 03:24:00	2016-01-01	3:24	Y	N	N
1308182	Moving Violation	310	W	M	24	2016-01-01 06:40:00	2016-01-01	6:40	Y	N	N
1308969	Moving Violation	Unknown	W	F	38	2016-01-01 06:45:00	2016-01-01	6:45	Y	N	N
1308181	Equipment Violation	830	H	M	18	2016-01-01 06:50:00	2016-01-01	6:50			
1308191	Moving Violation	230	W	M	25	2016-01-01 07:52:00	2016-01-01	7:52	N	N	N
1308183	Moving Violation	520	H	M	31	2016-01-01 08:15:00	2016-01-01	8:15	Y	N	N
1308187	Equipment Violation	510	H	M	31	2016-01-01 08:15:00	2016-01-01	8:15	Y	N	Y
1308186	Moving Violation	710	H	F	48	2016-01-01 08:21:00	2016-01-01	8:21	N	N	N
1308184	Equipment Violation	320	O	M	68	2016-01-01 08:25:00	2016-01-01	8:25	Y	N	N

DS: the real way

- What problem needs to be solved
  - in industry, solving a problem means providing value to the business
  - in research, even more complicated
  - Outcome:
    - a good question!
    - Easy to ask a question, isn't it?

https://towardsdatascience.com/how-to-ask-the-right-questions-as-a-data-scientist-913621907411

Right question

- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" --->

Right question

- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" ---> "The price will change"

Right question

- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" ---> "The price will change"
  - "What will my stock's sale price be next week?" ---> specific price!

Right question

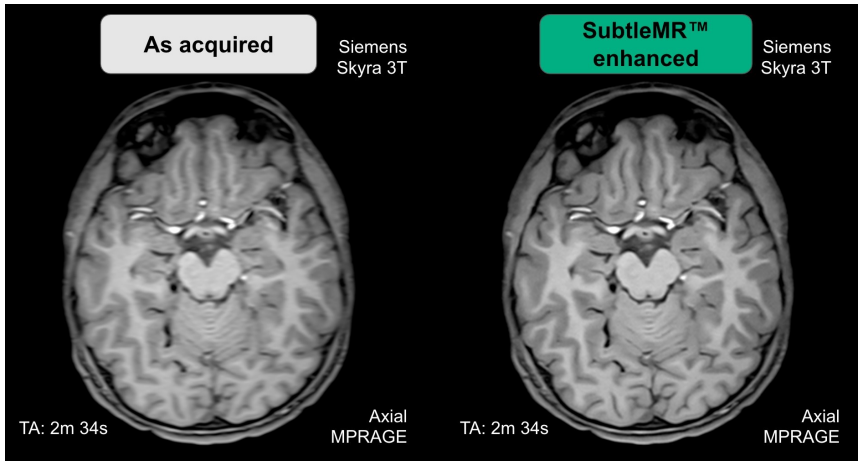
- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" ---> "The price will change"
  - "What will my stock's sale price be next week?" ---> specific price!
- Make sure your data can answer the question!
- Reformulate your question
  - insight from data
  - can they be generalized
  - can they be used for future prediction
- Questions we can answer now:
  - Is the police pulling over car at the right moment?
    - What time are cars usually pulled over?
    - What time are crashing usually happening?
    - Day of the week
    - Geographical area

https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-ask-a-question-you-can-answer-with-data

DS: the real way

- What problem needs to be solved
  - in industry, solving a problem means providing value to the business
  - in research, even more complicated
  - Outcome:
    - a good question!
    - Easy to ask a question, isn't it?

- Define success
  - What metric to use



https://towardsdatascience.com/how-to-ask-the-right-questions-as-a-data-scientist-913621907411

Back to our data

- Have you defined the right question for your Text Project?
  - Write me in Slack (during the lecture) what do you want to do, e.g.:
    - I want to compare the Wikipedia page of Michael Jordan and LeBron James. I want to answer the question: "Who was the best player?"

Back to our data

- Have you defined the right question for your Text Project?
  - Write me in Slack (during the lecture) what do you want to do, e.g.:
    - I want to compare the Wikipedia page of Michael Jordan and LeBron James. I want to answer the question: "Who was the best player?"
    - Well, let's pay a bit of attention, a better question would be: "Which is the most enthusiastic Wikipedia page between these two champions?"