

---

# **CSE 519: Data Science**

## **Steven Skiena**

### **Stony Brook University**

---

Lecture 1: Introduction to Data Science

---

# What is Data Science?

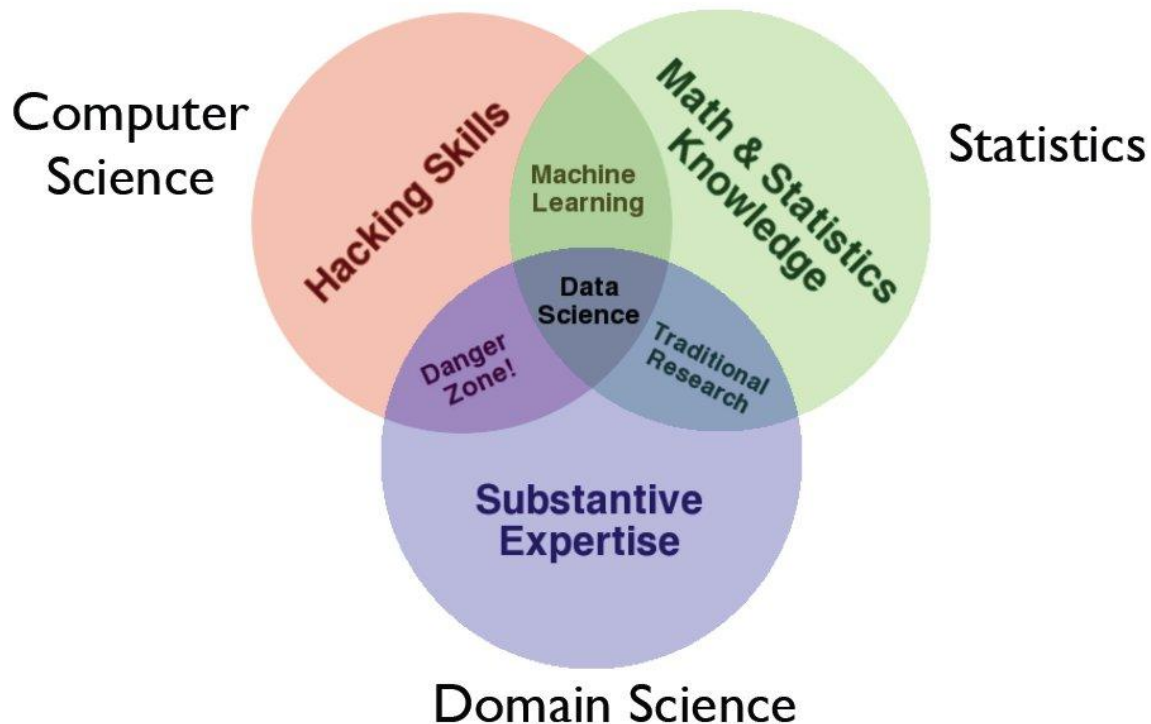
---

Like any emerging field, it isn't yet well defined, but incorporates elements of:

- Exploratory Data Analysis and Visualization
  - Machine Learning and Statistics
  - High-Performance Computing technologies for dealing with scale.
-

# Skill Sets for Data Science

---



# Appreciating Data

---

Computer Scientists do not naturally appreciate data: it's just stuff to run through a program.

The usual way to test algorithm performance is to run the implementation on “random data”.

But interesting data sets are a scarce resource, which requires hard work and imagination to obtain.

---

# Computer vs. Real Scientists (1)

---

- Scientists strive to understand the complicated and messy natural world, while computer scientists build their own clean and organized virtual worlds. Thus:
  - Nothing is ever completely true or false in science, while everything is either true or false in Computer Science / Mathematics.
-

# Computer vs. Real Scientists (2)

---

- Scientists are data-driven, while computer scientists are algorithm-driven.
  - Scientists obsess about discovering things, which computer scientists invent rather than discover.
  - Scientists are comfortable with the idea that data has errors; computer scientists are not.
-

# Genius vs. Wisdom

---

Software developers are hired to produce code.

Data Scientists are hired to produce insights.

Genius shows in finding the right answer!!!

Wisdom shows in avoiding the wrong answers.

Data science (like most things) benefits more from wisdom than from genius.

---

# Developing Wisdom

---

- Wisdom comes from experience.
- Wisdom comes from general knowledge.
- Wisdom comes from listening to others.
- Wisdom comes from humility, observing how often you have been wrong and why/how.

I seek pass on wisdom, through my experience on the difficulty of making good predictions.

---



# Developing Curiosity

---

- The good data scientist develops a curiosity about the domain/application they are working in.
  - They talk shop with the people whose data they are working on.
  - They read the newspaper every day, to get a broader perspective on the world.
-

# Asking Good Questions

---

Software developers are not encouraged to ask questions, but data scientists are:

- What exciting things might you be able to learn from a given data set?
  - What things do you/your people really want to know?
  - What data sets might get you there?
-

# Let's Practice Asking Questions!

---

Who, What, Where, When, and Why on the following datasets:

- [Baseball-reference.com](http://baseball-reference.com)
  - International Movie Database (IMDb)
  - Google ngrams
  - NYC taxi cab records
-

# Baseball-Reference.com: biosketch



play index **players** teams seasons managers leaders awards postseason boxes japan nlb minors draft

Mobile Site You Are Here > Home > Encyclopedia of Players > R Listing > Babe Ruth Statistics and

News: s-r blog:KBO Stats back to 1999 - Baseball-Reference.com

Babe Ruth Player Page » Batting Pitching Fielding Minors News Archive (1456) Bullpen Oracle



## Babe Ruth

Like 1,213 people like this.

+25 Recommend this

George Herman Ruth ([Babe](#), [The Bambino](#) or [The Sultan Of Swat](#))

**Positions:** Outfielder and Pitcher

**Bats:** Left, **Throws:** Left

**Height:** 6' 2", **Weight:** 215 lb.

**Born:** February 6, 1895 in Baltimore, MD

**High School:** St. Mary's HS (Baltimore, MD) (All Transactions)

**Debut:** July 11, 1914 (Age 19.155)

**Rookie Status:** Exceeded rookie limits during 1915 season [\*]

**Teams** (by GP): Yankees/RedSox/Braves 1914-1935

**Final Game:** May 30, 1935 (Age 40.113)

**Inducted into the Hall of Fame** by BBWAA as Player in 1936 (215/226 ballots). Induction ceremony in [View Babe Ruth Page](#) at the Baseball Hall of Fame (plaque, photos, videos).

**Died:** August 16, 1948 in New York, NY (Aged 53.192)

**Buried:** Gate of Heaven Cemetery, Hawthorne, NY

[View Player Bio](#) from the [SABR BioProject](#)

[About biographical information](#)



S-R: M

## Transactions

**July 9, 1914:** Purchased with [Ernie Shore](#) and [Ben Egan](#) by the [Boston Red Sox](#) from Baltimore (International) for more than \$25000. more than \$25000

**December 26, 1919:** Purchased by the [New York Yankees](#) from the [Boston Red Sox](#) for \$100,000.

**February 26, 1935:** Released by the [New York Yankees](#).

**February 26, 1935:** Signed as a Free Agent with the [Boston Braves](#).

The transaction information used here was obtained free of charge from and is copyrighted by [RetroSheet](#). We attempt to update transactions throughout the season.

## Salaries

Convert to YYYY \$5's

Salaries may not be complete (especially pre-1985) and may not include some earned bonuses

Year	Age	Team	Salary	ServTm(OnpDay)	Sources	Notes/Other Sources
1914	19	Boston Red Sox	\$2,500		? Bill James Historical Abstract	Annualized rate; came up late in season
1915	20	Boston Red Sox	\$3,500		? Bill James Historical Abstract	
1916	21	Boston Red Sox	\$3,500		? Contract at HOF	
1917	22	Boston Red Sox	\$3,500		? Contract at HOF	BJHA: \$5,000; Baseball Timeline \$7,000
1918	23	Boston Red Sox	\$9,000		? Allan Wood, 1918, at 183	Includes \$1,000 midseason raise, \$1,000 WS' bonus
1919	24	New York Yankees	\$10,000*		? Michael Haupert research of HOF contracts	Contract at HOF: 10000.00,
1920	25	New York Yankees	\$20,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 20000.00,
1921	26	New York Yankees	\$20,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 30000.00, Plus \$5K for '20 and '21 exhibitions, \$50/HR (\$9)m
1922	27	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 52000.00,
1923	28	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 52000.00,
1924	29	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 52000.00,
1925	30	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 52000.00,
1926	31	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 52000.00,
1927	32	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	5/23/27 AL letter: 70000.00,
1928	33	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	5/23/27 AL letter: 70000.00,
1929	34	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	5/23/27 AL letter: 70000.00,
1930	35	New York Yankees	\$70,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 80000.00,
1931	36	New York Yankees	\$70,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 80000.00,
1932	37	New York Yankees	\$70,000*		? Michael Haupert research of HOF contracts	M. Smelser, Life That Ruth Built, p. 441: 75000.00, Plus 25% of all exhibition-game profits
1933	38	New York Yankees	\$80,000*		? Michael Haupert research of HOF contracts	M. Smelser, Life That Ruth Built, p. 456: 52000.00, Plus 25% of revenue from in-season exhibitions
1934	39	New York Yankees	\$80,000*		? Michael Haupert research of HOF contracts	1/16/36 TSN, per government report: 36696.00, \$35,000 salary plus 25% of exhibition profits
1935	40	New York Yankees	\$75,000*		? Michael Haupert research of HOF contracts	Bill James Historical Abstract: 35000.00, Annualized rate; retired early in season
1936	41	New York Yankees	\$52,000*		? Michael Haupert research of HOF contracts	
1937	42	New York Yankees	\$35,000		? Michael Haupert research of HOF contracts	
Career to date (may be incomplete)			\$1,020,000			

# Statistical Record of Play

Summary  
statistics of each  
years batting,  
pitching, and  
fielding record,  
with teams and  
awards.

Babe Ruth Player Page

BattingPitchingFieldingMinorsNews Archive (1456)BullpenOracle

Fan EloRater

Line Details · Last updated Jun 3, 2014 9:17AM

All-Time Rank (among batters): #1. BABE RUTH... #2. Lou Gehrig... #3. Ted Williams... #4. Honus Wagner...

Vote

Standard BattingMore Stats

Glossary · Show Minors Stats · SHARE · Embed · CSV · PRE · LINK · ?

MinorsGame LogsSplitsHR LogFinders

Year	Age	Tm	Lg	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos	Awards	
1914	19	BOS	AL	5	10	10	1	2	1	0	0	0	0	0	0	4	.200	.200	.300	.500	49	3	0	0	0	0	0	/1		
1915	20	BOS	AL	42	103	92	16	29	10	1	4	20	0	0	9	23	.315	.376	.576	.952	188	53	0	2	1	0	0	1		
1916	21	BOS	AL	67	152	136	18	37	5	3	3	16	0	0	10	23	.272	.322	.419	.741	121	57	0	4	1	0	0	1		
1917	22	BOS	AL	52	142	123	14	40	6	3	2	14	0	0	12	18	.325	.385	.472	.857	162	58	0	7	1	0	0	1		
1918	23	BOS	AL	95	382	317	50	95	26	11	11	61	6	0	58	58	.300	.411	.555	.966	192	176	2	3	0	0	0	0	0	
1919	24	BOS	AL	130	543	432	103	139	34	12	29	113	7	0	101	58	.322	.456	.657	1.114	217	284	6	3	0	0	0	0	0	
1920	25	NY	AL	142	616	458	158	172	36	9	54	135	14	14	150	80	.376	.532	.847	1.379	255	388	3	5	0	0	0	0	0	
1921	26	NY	AL	152	693	540	177	204	44	16	59	168	17	13	145	81	.378	.512	.846	1.359	238	457	4	4	0	0	0	0	0	
1922	27	NY	AL	110	496	406	94	128	24	8	35	96	2	5	84	80	.315	.434	.672	1.106	182	273	1	4	0	0	0	0	0	
1923	28	NY	AL	152	697	522	151	205	45	13	41	130	17	21	170	93	.393	.545	.764	1.309	239	399	4	3	0	0	0	0	0	
1924	29	NY	AL	153	681	529	143	200	39	7	46	124	9	13	142	81	.378	.513	.739	1.252	220	391	4	6	0	0	0	0	0	
1925	30	NY	AL	98	426	359	61	104	12	2	25	67	2	4	59	68	.290	.393	.543	.936	137	195	2	6	0	0	0	0	0	
1926	31	NY	AL	152	652	495	139	184	30	5	47	153	11	9	144	76	.372	.516	.737	1.253	225	365	3	10	0	0	0	0	0	
1927	32	NY	AL	151	691	540	158	192	29	8	60	165	7	6	137	89	.356	.486	.772	1.258	225	417	0	14	0	0	0	0	0	
1928	33	NY	AL	154	684	536	163	173	29	8	54	146	4	5	137	87	.323	.463	.709	1.172	206	380	3	8	0	0	0	0	0	
1929	34	NY	AL	135	587	499	121	172	26	6	46	154	5	3	72	60	.345	.430	.697	1.128	193	348	3	13	0	0	0	0	0	
1930	35	NY	AL	145	676	518	150	186	28	9	49	153	10	10	136	61	.359	.493	.732	1.225	211	379	1	21	0	0	0	0	0	
1931	36	NY	AL	145	663	534	149	199	31	3	46	162	5	4	128	51	.373	.495	.700	1.195	218	374	1	0	0	0	0	0	0	
1932	37	NY	AL	133	589	457	120	156	13	5	41	137	2	2	130	62	.341	.489	.661	1.150	201	302	2	0	0	0	0	0	0	
1933	38	NY	AL	137	576	459	97	138	21	3	34	104	4	5	114	90	.301	.442	.582	1.023	176	267	2	0	0	0	0	0	0	
1934	39	NY	AL	125	471	365	78	105	17	4	22	84	1	3	104	63	.288	.448	.537	.985	160	196	2	0	0	0	0	0	0	
1935	40	BSN	NL	28	92	72	13	13	0	0	6	12	0	0	20	24	.181	.359	.431	.789	119	31	2	0	0	0	0	0	0	
22 Yrs					2503	10622	8399	2174	2873	506	136	714	2214	123	117	2062	1330	.342	.474	.690	1.164	206	5793	2	43	113				
162 Game Avg.					162	687	544	141	186	33	9	46	143	8	133	86	.342	.474	.690	1.164	206	375	3	7						
					G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos	Awards
NY (15 yrs)					2084	9198	7217	1959	2518	424	106	659	1978	110	117	1852	1122	.349	.484	.711	1.195	209	5131	35	94					
BOS (6 yrs)					391	1332	1110	202	342	82	30	49	224	13	0	190	184	.308	.413	.568	.981	190	631	8	19					
BSN (1 yr)					28	92	72	13	13	0	0	6	12	0	0	20	24	.181	.359	.431	.789	119	31	2	0	0				
AL (21 yrs)					2475	10530	8327	2161	2860	506	136	708	2202	123	117	2042	1306	.343	.475	.692	1.167	207	5762	43	113					
NL (1 yr)					28	92	72	13	13	0	0	6	12	0	0	20	24	.181	.359	.431	.789	119	31	2	0	0				

# Baseball Questions

---

- How to best measure individual player's skill, value or performance?
  - How fair do trades between teams work out?
  - What is the trajectory of player's performances as they mature and age?
  - To what extent does batting performance correlate with the position played?
-

# Demographic Questions

---

- Do left-handed people have shorter lifespans than right-handers?
  - How often do people return to where they were born?
  - Do player salaries reflect past, present, or future performance?
  - Are heights and weights increasing in the population?
-



# IMDb: Movie Data

All

[Movies, TV & Showtimes](#) [Celebs, Events & Photos](#) [News & Community](#) [Watchlist](#)



LIBERTY FILMS - A Frank CAPRA'S  
It's A WONDERFUL LIFE  
Starring James STEWART and Donna REED  
LIONEL BARRYMORE - THOMAS MITCHELL - HENRY TRAVERS -  
BURLAP HEND - ARPA - DORIS - FRANK SPOTER - ELUDA - GUARANE -  
Produced by FRANK CAPRA

**It's a Wonderful Life** (1946)  **Top 5000**

**Approved** 130 min - Drama | Family | Fantasy -  
7 January 1947 (USA)

**Your rating:** ★★★★★★ ★★ -/10  
Ratings: **8.7**/10 from 202,743 users  
Reviews: 632 user | 187 critic

An angel helps a compassionate but despairingly frustrated businessman by showing what life would have been like if he never existed.

**Director:** Frank Capra  
**Writers:** Frances Goodrich (screenplay), Albert Hackett (screenplay), 4 more credits »  
**Stars:** James Stewart, Donna Reed, Lionel Barrymore | See full cast and crew »

[+ Watchlist](#) [Watch Trailer](#) [Share...](#)

[More at IMDbPro »](#)

## Details

[Edit](#)

**Country:** USA

**Language:** English

**Release Date:** 7 January 1947 (USA) [See more »](#)

**Also Known As:** The Greatest Gift [See more »](#)

**Filming Locations:** California, USA [See more »](#)

## Box Office

**Budget:** \$3,180,000 (estimated)

**Opening Weekend:** £49,845 (UK) (19 December 2008)

**Gross:** £682,222 (UK) (24 December 2010)

[See more »](#)

## Company Credits

**Production Co:** Liberty Films (II) [See more »](#)

[Show detailed company contact information on IMDbPro »](#)

## Technical Specs

**Runtime:** 130 min | 118 min (DVD edition)

**Sound Mix:** Mono (RCA Sound System)

**Color:** Color (colorized) | Black and White

**Aspect Ratio:** 1.37 : 1

[See full technical specs »](#)



# IMDb: Actor Data



## James Stewart (I) (1908–1997)

[Actor](#) | [Soundtrack](#) | [Director](#)

James Maitland Stewart was born on 20 May 1908 in Indiana, Pennsylvania, where his father owned a hardware store. He was educated at a local prep school, Mercersburg Academy, where he was a keen athlete (football and track), musician (singing and accordion playing), and sometime actor. In 1929 he won a place at Princeton, where he studied ... [See full bio »](#)

**Born:** James Maitland Stewart  
May 20, 1908 in Indiana, Pennsylvania, USA

**Died:** July 2, 1997 (age 89) in Los Angeles, California, USA



[230 photos](#) | [42 videos](#) | [1180 news articles](#) »

**Won 1 Oscar.** Another 25 wins & 19 nominations. [See more awards](#) »

## Cast

[Edit](#)

Cast overview, first billed only:

	James Stewart	...	George Bailey
	Donna Reed	...	Mary Hatch
	Lionel Barrymore	...	Mr. Potter
	Thomas Mitchell	...	Uncle Billy
	Henry Travers	...	Clarence
	Beulah Bondi	...	Mrs. Bailey
	Frank Faylen	...	Ernie
	Ward Bond	...	Bert
	Gloria Grahame	...	Violet
	H.B. Warner	...	Mr. Gower

# Movie Questions

---

- Can we predict how well people will like a movie? What about its gross?
  - What does the social network of actors look like? (Six degrees of Kevin Bacon)
  - What is the age distribution of actors and actresses in film?
  - Do stars live longer or shorter lives than the bit players or public?
-

# Google Ngrams

---

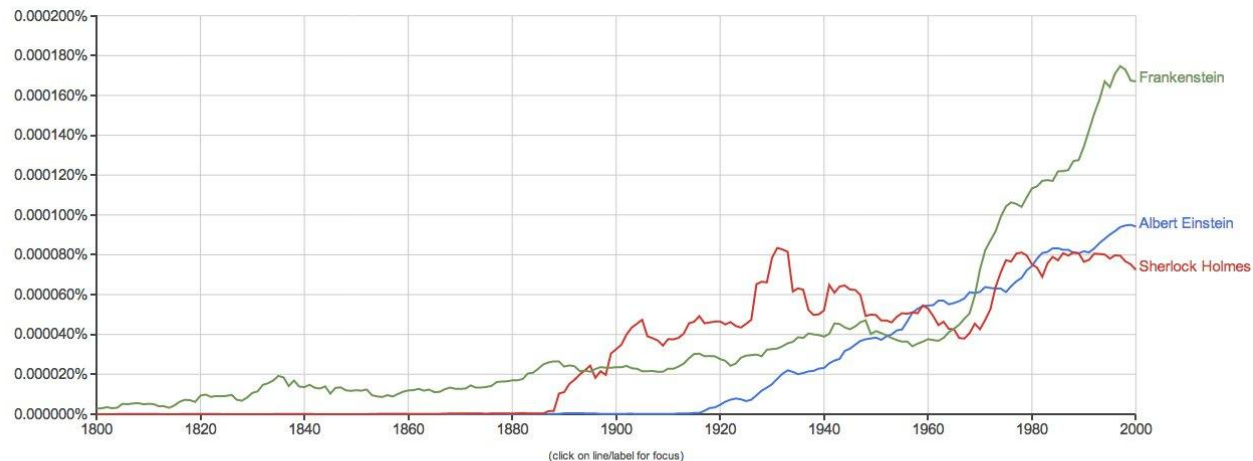
- Presents an annual time series of the frequency of every “popular” word/phrase with 1 to 5 words occurs in scanned books.
  - ‘Popular’ means appears >40 times in total.
  - Google has scanned about 15% of all books ever published, making this resource quite comprehensive.
-

# Google Ngram Viewer

## Google books Ngram Viewer

Graph these comma-separated phrases:  ☐ case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)



Run your own experiment! Raw data is available for download [here](#).

# Ngram Questions

---

- How has the amount of cursing changed over time?
  - What is the lifespan of fame and technologies? Is it increasing/decreasing?
  - How often do new words emerge? Do they stay in common usage?
  - What words are associated with other words, i.e. can you build a language model?
-

# NYC Taxi Cab Data

- Gives driver/owner, pickup/dropoff location, and fare data for every taxi trip taken.
- Data obtained from NYC via Freedom of Information Act Request (FOA)

4													
5	Trip data, 2013 ->												
6													
7	medallion	hack_license	vendor_id	rate_code	pickup_datetime	dropoff_datetime	passenger_count	trip_time	trip_distance	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
8	89D227B655E5C82AEC	BA96DE419E7116	CMT	1	1/1/13 15:11	1/1/13 15:18	4	382	1	-73.978165	40.757977	-73.989838	40.751171
9	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1	1/6/13 0:18	1/6/13 0:22	1	259	1.5	-74.006683	40.731781	-73.994499	40.75066
10	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1	1/5/13 18:49	1/5/13 18:54	1	282	1.1	-74.004707	40.73777	-74.009834	40.726002
11	...												
12													
13													
14	Fare data, 2013 ->												
15													
16	medallion	hack_license	vendor_id	pickup_datetime	fare_amount	surcharge	mta_tax	tip_amount	tolls_amount	total_amount			
17	89D227B655E5C82AEC	BA96DE419E7116	CMT	1/1/13 15:11	6.5	0	0.5	0	0	7			
18	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1/6/13 0:18	6	0.5	0.5	0	0	7			
19	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1/5/13 18:49	5.5	1	0.5	0	0	7			

# Taxicab Questions

---

- How much do drivers make each night?
  - How far do they travel?
  - How much slower is traffic during rush hour?
  - Where are people traveling to/from at different times of the day?
  - Do faster drivers get tipped better?
  - Where should drivers go to pick up their next fare?
-