LEAH GAETA

DATA SCIENCE FINAL PROJECT PROPOSALS

NCAA APR PREDICTOR

- Problem Statement:
  - Using NCAA eligibility, retention, and previous Academic Progress Rate (APR) scores from 2004 - 2014, predict future postseason eligibility (at least 930 four-year average APR, or 940 over two most recent years)
  - Predict patterns based on sport, explore differences between men's and women's teams
- Data:
  - https://www.kaggle.com/ncaa/academic-scores
  - Calculate APR: http://www.ncaa.org/aboutresources/research/academic-progress-rate-explained
- Hypothesis:
  - Given data will allow one to predict & identify teams at risk of losing postseason eligibility

SCHOOL_ID	SCHOOL_NAI	SCHOOL_TYP	ACADEMIC_Y	SPORT_CODE	SPORT_NAM	NCAA_DIVISI	NCAA_SUBDI	NCAA_CONFI	FOURYEAR_A	FOURYEAR_S	FOURYEAR_E	FOURYEAR_F	2014_ATHLE	2014_SCORE	2014_ELIGIBI	2014_RETEN
100654	Alabama A&I	0	2014	1	Baseball	1	2	Southwester	80	931	0.902	0.9536	21	976	0.9762	0.9762
100654	Alabama A&I	0	2014	4	Football	1	2	Southwester	321	932	0.8861	0.9613	78	905	0.8108	0.9797
100654	Alabama A&I	0	2014	2	Men's Basket	1	2	Southwester	43	964	0.9177	0.988	12	958	0.9167	1
100654	Alabama A&I	0	2014	6	Men's Golf	1	2	Southwester	22	898	0.7949	0.8462	4	1000	0.875	1
100654	Alabama A&I	0	2014	13	Men's Tennis	1	2	Southwester	12	988	0.913	1	-99	-99	-99	-99
100654	Alabama A&I	0	2014	14	Men's Track,	1	2	Southwester	62	932	0.8833	0.9829	18	910	0.8824	0.9394
100654	Alabama A&I	0	2014	15	Men's Track,	1	2	Southwester	59	939	0.8879	0.9913	17	909	0.8485	0.9697
100654	Alabama A&I	0	2014	19	Women's Bas	1	2	Southwester	53	990	0.9903	0.9899	14	980	1	0.9565
100654	Alabama A&I	0	2014	20	Women's Bo	1	2	Southwester	24	1000	1	1	5	1000	1	1
100654	Alabama A&I	0	2014	21	Women's Cro	1	2	Southwester	33	985	0.9539	1	7	1000	1	1

MLB BALLPARK ATTENDANCE PREDICTOR

- Problem Statement:
  - Using team offensive & defensive data from the complete history of major league baseball stats collected from 1960 - 2015, create a model to predict MLB Ballpark Attendance for each team
  - Does play affect attendance?
  - Offensive Predictor Variables: R, HR, H, SOA, SB, etc.
  - Defensive Predictor Variables: ERA, K, E, SO, RA, etc.
- Data:
  - https://www.kaggle.com/seanlahman/the-history-of-baseball
- Hypothesis:
  - Given data will allow one to predict ballpark attendance based on team performance

year	league_id	team_id	g	ghome	w	l	div_win	wc_win	lg_win	ws_win	r	ab	h	double	triple	hr	
2014	NL	ARI	162	81	64	98	N	N	N	N	615	5552	1379	259	47	118	
2014	NL	ATL	162	81	79	83	N	N	N	N	573	5468	1316	240	22	123	
2014	AL	BAL	162	81	96	66	Y	N	N	N	705	5596	1434	264	16	211	
2014	AL	BOS	162	81	71	91	N	N	N	N	634	5551	1355	282	20	123	
2014	AL	CHA	162	81	73	89	N	N	N	N	660	5543	1400	279	32	155	
2014	NL	CHN	162	81	73	89	N	N	N	N	614	5508	1315	270	31	157	
2014	NL	CIN	162	81	76	86	N	N	N	N	595	5395	1282	254	20	131	
2014	AL	CLE	162	81	85	77	N	N	N	N	669	5575	1411	284	23	142	
2014	NL	COL	162	81	66	96	N	N	N	N	755	5612	1551	307	41	186	

MLB TEAM WIN PERCENTAGE PREDICTOR

- Problem Statement:
  - Using MLB player salary data from the complete history of major league baseball stats collected from 1985 - 2015, create a model to predict team win percentage
  - Are the biggest spenders the biggest winners?
- Data:
  - https://www.kaggle.com/seanlahman/the-history-of-baseball
- Hypothesis:
  - Given data will allow one to predict win percentage based on how much a team spends on its players

	team_id	league_id	player_id	salary
1985	ATL	NL	barkele01	870000
1985	ATL	NL	bedrost01	550000
1985	ATL	NL	benedbr01	545000
1985	ATL	NL	campri01	633333
1985	ATL	NL	ceronri01	625000
1985	ATL	NL	chambch01	800000
1985	ATL	NL	dedmoje01	150000
1985	ATL	NL	forstte01	483333
1985	ATL	NL	garbege01	772000

year	league_id	team_id	g	ghome	w	l	div_win	wc_win	lg_win	ws_win
1985	NL	ATL	162	81	66	96	N		N	N
1985	AL	BAL	161	81	83	78	N		N	N
1985	AL	BOS	163	81	81	81	N		N	N
1985	AL	CAL	162	79	90	72	N		N	N
1985	AL	CHA	163	81	85	77	N		N	N
1985	NL	CHN	162	81	77	84	N		N	N
1985	NL	CIN	162	81	89	72	N		N	N
1985	AL	CLE	162	81	60	102	N		N	N
1985	AL	DET	161	81	84	77	N		N	N

THOUGHTS?