데이터 분석가
직무 포트폴리오

심보람

목차 / 빅데이터 공모전 / 프로젝트 / 졸업논문 / 컨설팅

프로젝트
- 자전거 공유 서비스 수익률 예측

- 프로젝트 소개
- 데이터 분석
- 분석기법 및 분석결과

통계 컨설팅 프로젝트
- 코티솔 관계

- 컨설팅 프로젝트 소개
- 데이터 수집 및 분석
- 분석 기법 및 분석결과

빅데이터 공모전

- 공모전 소개
- 데이터 수집 및 분석
- 분석기법 및 분석 결과

졸업논문

- 졸업논문 소개
- 데이터 수집 및 분석
- 분석기법 및 분석 결과

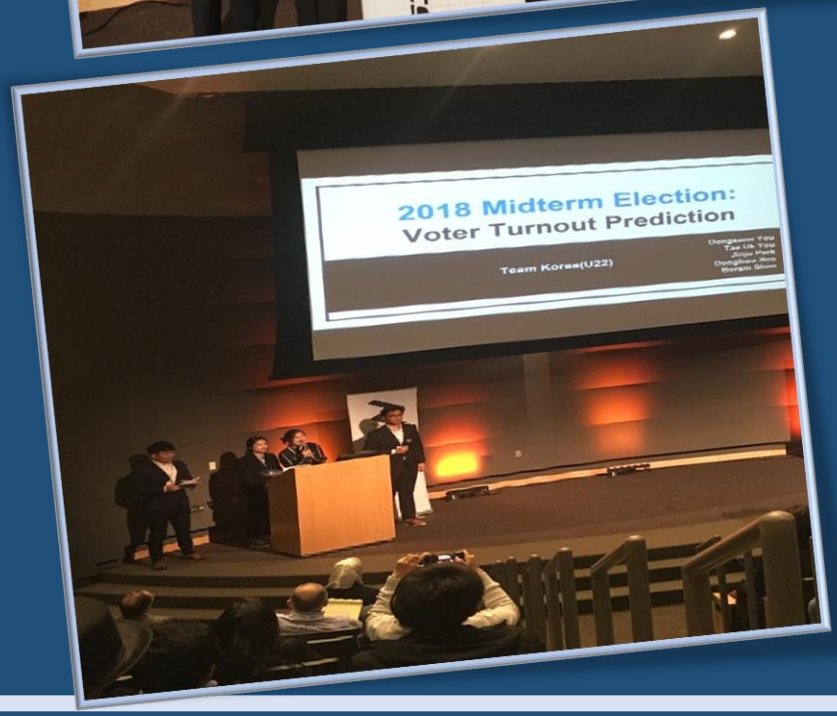
MinneMUDAC 2018 빅데이터 공모전

MinneMUDAC 2018 공모전 소개

- 미국 미네소타주 분석커뮤니티(MinneAnalysics)와 미국 중서부 대학 데이터 공모전(MUDAC) 주관 빅데이터 공모전

- 주제: 2018년도 미국 중간선거 투표자수를 미네소타 주 8개 선거구 별로 예측

- 5명의 팀원과 어드바이저로 모신 통계학과 교수님 (Prof. Glen D Meeden)으로 팀 구성



데이터 수집

미국 기상청

- 선거 다일 평균 기온과 강수량

CNN

- 후보자 간 경쟁률 차이

미국 통계청

- 인종 (White, Hispanic, Other)
- 성별(M / F)
- 나이: 18세 이상 (4단계 분류)
- 교육수준 (4단계 분류)
- 소득수준 등 23개의 변수

미네소타 주 선거 관리국

- 8개의 선거구별 투표자 수

데이터 분석

- 역대 최고 투표율이 예상되는 2018년 중간선거
- 대선보다 상대적으로 낮은 투표율
- 누락된 데이터로 2006년 이후의 데이터 수집
- 데이터 부족으로 중간선거 (2006, 2010, 2014)와 대선(2008, 2012, 2016) 데이터를 분석
  - 6년 동안의 미네소타 8개 주의 데이터, 즉 48개의 데이터만이 사용

대선 포함 선거 데이터와
중간선거만 포함한 선거 데이터
선택 필요

중간선거 투표율 70% 넘는 '역대급' 전망

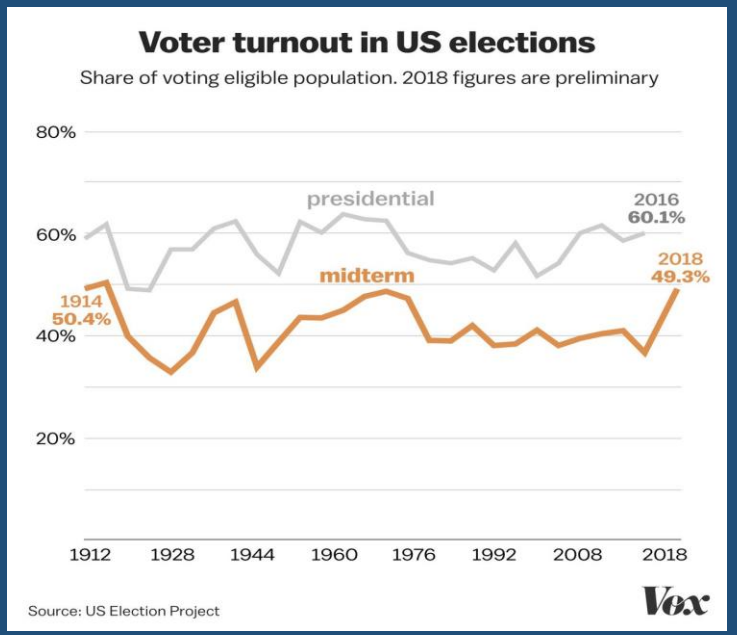
임상환 기자

역대 최고 투표율 예상, '트럼프-반트럼프' 중간선거 승자는?

SBS CNBC

美 중간선거 투표율 60% 넘을 듯..오늘 정오 투표 마무리

김영교 기자 입력 2018.11.07 09:57 댓글 0개



모델링 및 투표자 수 예측

년도	선거
2006	중간선거
2010	중간선거
2014	중간선거
2018	중간선거

2006년, 2010년 중간선거 데이터를 활용하여 모델링 후 2014년 투표자 수 예측 및 실제 2014년 투표자 수 비교

년도	선거
2006	중간선거
2008	대통령 선거
2010	중간선거
2012	대통령 선거
2014	중간선거
2016	대통령 선거
2018	중간선거

2006, 2008, 2010, 2012, 2014년 대선과 중간선거 데이터를 활용하여 모델링 후 2016년 투표자 수 예측 및 실제 2016년 투표자 수 비교

실제 투표자 수와 예측값 비교

year	actual voter turnsout	model (predicted/actual) PLS	ABS(value-1)	if(2014>2016,1,0)
2014	229564	0.96981626	0.03018374	0
2014	248549	1.020228204	0.020228204	0
2014	273488	1.017266205	0.017266205	0
2014	246088	1.080331833	0.080331833	0
2014	240709	1.07426353	0.07426353	0
2014	240697	0.876913713	0.123086287	1
2014	244791	0.924059708	0.075940292	0
2014	268680	0.888349338	0.111650662	0
		PCR		
2016	346854	1.222938758	0.222938758	
2016	384539	1.174476971	0.174476971	
2016	405198	1.151753217	0.151753217	
2016	370000	1.135786216	0.135786216	
2016	376895	1.197326046	0.197326046	
2016	376481	1.104859741	0.104859741	
2016	342584	1.189639621	0.189639621	
2016	365730	1.225801001	0.225801001	

- 비율에서 1을 뺀 절대값이 작을수록 좋음
- 2014년 모델의 절대값이 2016년보다 작은 경우가 더 많음

결과: 중간선거 데이터 사용

분석 기법

- **5 종류의 분석 방법을 활용**
  - ✓ Principle Component Regression (PCR): 데이터의 분산과 과다 적합을 줄여 효율성을 증가
  - ✓ Partial Least Square (PLS): 독립변수와 종속변수를 구분하여 영향력 있는 변수 선택
  - ✓ Decision Tree: 데이터 탐색 및 시각화
  - ✓ Bagging: Bootstrapping을 통하여 분산을 낮추고 다중 트리로 높은 예측 정확도 증가
  - ✓ Random Forest: 다중 트리들 간의 연관성을 줄이기 위하여 전체 변수 중 랜덤하게 변수를 선택하여 다중 트리들이 같은 결과를 가져올 확률을 감소
- **K - fold Cross Validation**
  - ✓ 예측할 년도의 데이터 이전 년도 데이터를 활용해 모델링
  - ✓ 데이터를 10개의 fold로 나누고 각 fold에 train set과 validation set으로 나눈 후 10번 반복

모델 분석

- 중간선거 (2006, 2010, 2014) 데이터를 활용
- 5 종류의 분석기법을 활용하여 train set과 validation set으로 예측 오차율 도출 및 각 분석 예측 오차율 비교
- PCR 모델의 Mean Square Error가 가장 낮음

	MSE
PCR	387.8501
PLS	682.1454
Decision tree	4289.543
Prune	4531.051
Bagging	1792.053
Random Forest	1635.299

중간선거 데이터를 바탕으로 PCR을 활용하여 2018년도 중간 선거 투표자 수를 예측하는 것이 가장 정확할 수 있음을 알 수 있음

분석 결과

2018 중간선거 투표 결과

예측 투표자 수	
District 1	280,245
District 2	288,797
District 3	330,244
District 4	296,466
District 5	315,381
District 6	291,445
District 7	272,583
District 8	293,737

실제 투표자 수	
District 1	291,661
District 2	337,968
District 3	363,949
District 4	328,614
District 5	343,358
District 6	315,726
District 7	281,509
District 8	314,211

Serendipitous Discovery 수상

MinneMUDAC 2018 Outcomes

Round #1 Outcomes			
Rank	Novice	Undergraduate	Graduate
1	N13	U22	G9
2	N8	U1	G16
3	N20	U28	G18
4	N18	U11	G17
5	N16	U29	G7
6	N7	U14	G12
7	N12	U2	G1
26		U15	
27		U33	
28		U30	
29		U5	

Prediction Outcomes		
Rank	Undergraduate	Graduate
1	U6	G7
2	U14	G3
3	U22	G5
4	U1	G12
5	U15	G16
6	U19	G17
7	U26	G15
26	U27	
27	U5	
28	U30	
29	U13	

Round #2 Outcomes - Final Awards			
	Division		
Award	Novice	Undergraduate	Graduate
Overall	N13	U11	G16
Analytical Acumen	N8	U28	G18
Serendipitous Discovery	N20	U22	G9
Best Prediction		U6	G7
Honorable Mention Best Prediction		U14	G3

Undergraduate Division

First Place Overall: St. Olaf College Mathematics – *Junior Varsity Stats*

Faculty Advisor: Paul Roback

Analytical Acumen: University of Iowa Management Science – *TippieAnalytics Team 1*

Faculty Advisor: Michael Altemeier

Serendipitous Discovery: University of Minnesota College of Liberal Arts Economics, Statistics – *Team Korea*

Faculty Advisor: Glen Meeden

Best Prediction: University of Minnesota Duluth Department of Management – *Duluth's Data Dragons*

Faculty Advisor: Nik Hassan

Honorable Mention Prediction: St Olaf College Mathematics, Statistics, and Computer Science – *Model Citizens*

Faculty Advisor: Paul Roback and Matthew Richey

TeamID: U22

	[4:Superior to 1:Emerging]	
Item	Your Average Score	Division Average Score
Accuracy of Prediction	3.77	2.58
Completeness and breadth in answering questions	3.6	2.99
Appropriateness of analytical methods used	3.4	2.69
Creativity and Innovation	2.8	2.77
Communication of Results	3.4	2.84
Data Preparation	3.4	2.84
Team Synergy	3.4	3.13

프로젝트
런던 자전거 공유 서비스 수익률 예측

프로젝트 소개

- 런던의 자전거 공유 서비스를 운영하는 회사의 손실과 이익을 예측할 수 있는 모델 구축
- 회사에 이익을 가져다 주는 변수를 이해
- 하루 동안 공유되는 자전거의 수가 2만 대보다 적으면 회사는 손실 , 최소 2만대 이상의 자전거를 공유해야 서비스가 수익성이 높아진다고 가정
- 2015년부터 2016년까지 런던에서 매일 공유되는 자전거 수에 대한 정보 바탕

데이터 분석

- 예측하고자 하는 반응 변수를 N_bikes를 사용하여 구성 (손실=0, 이익=1)

  Profit=ifelse(bike_final$N_bikes<20000,0,1)

- 연속 변수 중 누락된 값을 포함하는 경우 해당 변수의 평균을 적용, 범주형 변수 중 누락된 값을 포함하는 경우 누락된 관측치 제외
- 결측자료(Missing data)를 폐기(discard), 비결측자료들로 회귀분석 후 대체(impute)하는 두가지의 방식으로 결측자료 분석

	date	N_bikes	temperature	feels_like	humidity	wind_speed	holiday	weekend	season
1	2016-09-27	35751	17.104167	17.1041667	75.87500	15.625000	0	0	2
2	2016-01-31	NA	10.041667	8.7708333	79.16667	22.500000	0	1	3
3	2016-06-04	NA	15.145833	15.1458333	NA	8.604167	0	1	1
4	2016-05-21	26373	15.937500	15.9375000	74.06250	19.250000	0	1	0
5	2016-07-06	42412	17.729167	17.6250000	53.54167	9.604167	0	0	1
6	2015-04-01	23539	8.500000	4.9375000	59.85417	31.708333	0	0	0
7	2016-08-01	29749	16.812500	16.8125000	67.87500	12.583333	0	0	1
8	2016-01-18	NA	2.937500	0.1250000	75.37500	11.437500	0	0	3
9	2015-03-31	22602	10.850000	9.7000000	55.15000	41.900000	0	0	0
10	2016-03-26	10053	10.815789	9.5000000	79.68421	22.578947	0	1	0
	date	N_bikes	temperature	feels_like	humidity	wind_speed	holiday	weekend	season
486	2016-09-04	27512	18.854167	18.85416667	69.85417	23.395833	0	1	2
487	2015-10-28	23881	13.583333	13.58333333	84.77083	10.875000	0	0	2
488	2015-02-17	25064	5.854167	3.47916667	68.70833	12.187500	0	0	3
489	2016-02-06	13126	11.041667	9.70833333	75.75000	31.083333	0	1	3
490	2015-12-12	15815	11.675000	11.20000000	73.07500	23.375000	0	1	3
491	2015-03-01	15641	9.187500	6.41666667	NA	29.583333	0	1	0
492	2015-02-03	NA	2.166667	-1.04166667	85.50000	11.833333	0	0	3
493	2015-12-30	12301	12.583333	12.54166667	69.37500	25.083333	0	0	3
494	2016-01-17	NA	2.958333	0.75000000	80.31250	9.062500	0	1	3
495	2015-06-18	38687	17.875000	17.87500000	57.25000	18.083333	0	0	1
496	2016-12-18	14970	8.145833	7.20833333	90.72917	6.520833	0	1	3
497	2015-12-15	NA	11.458333	11.12500000	92.25000	10.708333	0	0	3
498	2015-11-02	28033	10.500000	10.04166667	95.00000	9.166667	0	0	2
499	2015-04-26	NA	9.645833	8.29166667	83.33333	11.187500	0	1	0
500	2016-08-31	38989	19.812500	19.81250000	63.12500	15.333333	0	0	1

분석 기법 및 결과

- **4 종류의 분석 방법을 활용**
  - ✓ Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest
- **K-fold Cross Validation**
  - ✓ 데이터를 10개의 fold로 나누고 각 fold에 train set과 validation set으로 나눈 후 10번 반복
- **4 종류의 분석기법을 활용하여 train set으로 모델링 후 validation set으로 Test Error Rate, AUC 도출**
- **두 가지 결측자료 처리방법을 이용하여 나온 결과 비교**

Data 1 | Data2	Logistic Regression		KNN K=5		Classification Tree		Pruned Tree		Random Forest	
AUC	0.9295	0.5737	0.8714	0.9051	0.7527	0.8283	0.7116	0.7979	0.7429	0.8566
Test Error	0.08139	0.09	0.08139	0.12	0.1279	0.1	0.1163	0.1	0.093	0.08

Data1: 결측 자료 폐기(discard), Data2: 결측 자료 대체(impute)

결측 자료 대체 데이터에서, KNN모델의 K=15 일 때 AUC= 0.9288으로 1에 매우 가깝고, Test Error Rate= 0.06으로 매우 낮음 결국 Data 2를 바탕으로한 KNN 모델 활용이 손익 예측하는 가장 적합한 모델

졸업논문
리그오브레전드 승/패 요소 예측 분석

졸업논문 소개

- 현재 전세계적으로 매달 1억 명 이상의 사용자가 게임하는 리그오브레전드 승리 요소를 분석
- 2016년부터 2019년 까지 KT 롤스터즈의 League of Legends Championship Korean (LCK) 정보 바탕
- Limitation
  - ✓ 각각 능력치 다른 오브젝트들의 파밍 제외
  - ✓ 숫자로 표시 할 수 없는 주요 전략, 챔피언 상성들은 제외

데이터 수집

YouTube

- 리그오브레전드 분석 영상
- 리그오브레전드 게임 영상
- 리그오브레전드 게임 결과

LOL INVEN

- 2016년 ~ 2019년 게임 전적
  각 19개의 경기를 분석 (79번의 경기)

KT 롤스터

- 선수단 정보

OP.GG

- 챔피언 분석
- 선수들의 일반게임 분석
- 롤인벤에서 부족한 정보

데이터 분석

- 예측하고자 하는 반응 변수를 Tower를 사용하여 구성
- 15개의 타워 중 타워를 더 많이 깨는 것이 승리 요인이라 가정
- 타워, Kill/Death/Assist (KDA), 게임시간, 획득골드 등 데이터 정규화
- Akaike Information Criteria (AIC) 결과를 바탕으로 모델링 구축

분석 기법 및 결과

- 순서형 로지스틱 모형을 활용하여 타워 파괴 개수에 따른 랭크를 부여(High, Mid, Low)
- Akaike Information Criteria (AIC) 결과를 바탕으로 최종 모델링 구축
- 두 결과 AIC 값이 비슷하지만 p-value 값이 낮은 Kill만을 가지고 모델링
- Kill이 많을 수록 타워를 더 많이 파괴할 수 있음을 알 수 있음

  = (Kill이 1 증가할 때 OR은 1.56만큼 증가)

Coefficient	j= 1	j=2	P value
Intercept	1.834 (se= 1.645)	3.665 (se= 1.734)	0.2648(j=1) 0.0345 (j=2)
Kill	0.373, (se= 0.091)		0.00004
Assist	0.124 (se= 0.06565)		0.059
Time	-0.044 (se= 0.04873)		0.362
Residual deviance	87.123		
AIC	97.123		

Coefficient	j= 1	j=2	P value
Intercept	3.006 (se= 0.729)	4.728 (se= 0.893)	0.00001 (j=1). 0 (j=2)
Kill	0.445 (se= 0.0792)		0
Residual deviance	91.718		
AIC	97.718		

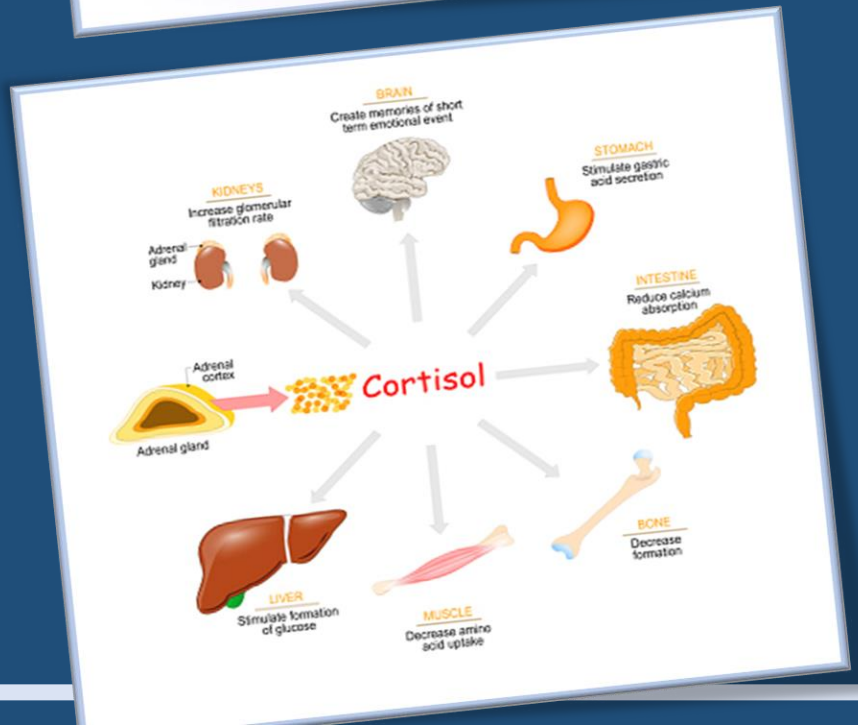
	j=1 (Low)		j=2 (Medium | High)	
	Less Kill	High Kill	Less Kill	High Kill
Odds Ratio	1.560 or exp (0.445)			

컨설팅
유아와 산모의 코티솔 레벨 상관관계

컨설팅 소개

- 공중 보건학 석사 전공자들의 산모와 유아 *코티솔 수준이 상관 관계가 있는지를 결정하는 것

- 통계에 대한 지식이 전무한 연구자들을 위해 알기 쉽게 용어 설명 및 데이터 분석제공



*코티솔: 부신 피질에서 분비되는 스테로이드 호르몬이며 급성 스트레스에 반응해 분비되는 물질

데이터 수집

- Islands for School Projects 웹사이트의 가상의 인구를 대상으로 설문 조사 후 데이터 수집
- 산모와 유아 코티솔 레벨 자료를 수집하여 산모의, 나이, 모유수유 여부, 스트레스 레벨에 의한 상관 관계와 영아의 성별, 나이 등이 어떤 영향을 미칠지 분석


```
'data.frame':   56 obs. of  18 variables:
 $ Family..                : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Mother.s.Name           : Factor w/ 44 levels "","Ai Regan ",..: 6 27 14 24 12 26 20 38 44 33 ...
 $ Infant.s.Name           : Factor w/ 44 levels "","Akira Suzuki ",..: 19 37 10 38 16 15 13 33 41 7 ...
 $ Sibiling.s.Name         : Factor w/ 44 levels "","Aimon Gagnon",..: 14 18 11 24 23 2 35 30 19 13 ...
 $ House..                 : int  530 85 466 19 1134 982 1449 1162 263 94 ...
 $ Village                 : Factor w/ 24 levels "","Akkeshi","Biruwa",..: 15 15 15 5 5 5 5 5 20 14 ...
 $ Island..North..South..East.: Factor w/ 6 levels "","East","East ",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ Mother.s.Age..years.    : int  43 27 40 21 22 27 25 29 24 29 ...
 $ Infant.s.DOB            : Factor w/ 25 levels "","05/285","07/285",..: 17 15 10 25 21 13 12 12 9 22 ...
 $ Infant.s.Age..days.     : int  17 19 24 5 13 21 22 22 25 12 ...
 $ Sibiling.s.Age..years.  : int  9 9 7 2 2 2 3 9 2 2 ...
 $ Infant.s.Sex            : Factor w/ 3 levels "","F","M": 2 3 2 3 2 2 3 3 2 2 ...
 $ Sibling.s.Sex           : Factor w/ 3 levels "","F","M": 2 3 3 2 2 3 3 3 3 3 ...
 $ Mother.s.Cortisol       : num  6.94 7.3 6.8 7.53 6.19 7.29 7.45 6.05 8.01 7.16 ...
 $ Infant.s.Cortisol       : num  4.44 4.46 4.5 3.77 3.76 4.31 4.87 4.12 5.48 3.47 ...
 $ Sibling.s.Cortisol      : num  6.07 5.63 5.94 5.75 5.26 5.94 5.96 5.28 5.89 5.75 ...
 $ Mother.s.Stress.Ranking : num  3 3.5 3 2 2.5 2 2.5 3 5 3 ...
 $ Mother.Breastfeeding..Y.N. : int  0 0 0 1 0 0 0 0 0 0 ...
> View(idata)
```

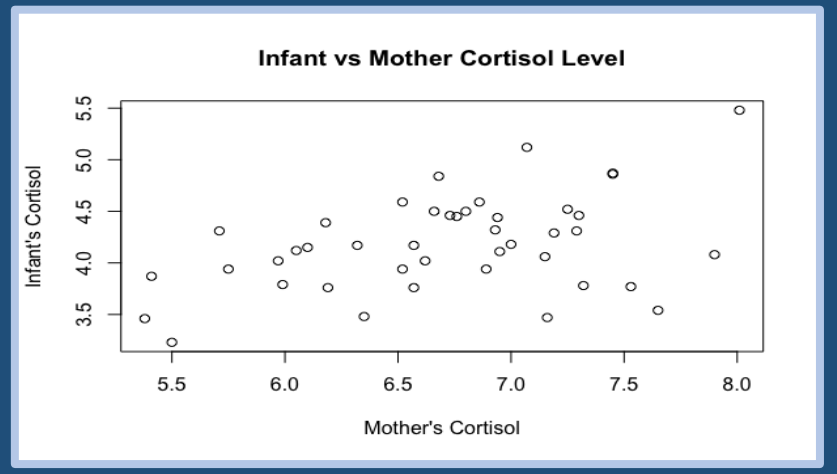
데이터 분석

- 그래프에 의하면 초기 가정과 다르게 산모와 유아보다 산모와 형제/자매 사이에서 강한 상관관계가 나타남
- 영아와 형제/자매는 산모의 영향을 받기 때문에, 모유 수유도 영향을 미침
- Limitation: 모유수유 기준을 포함하면 결과가 달라질 수 있는데, 주어진 데이터에서는 모유수유를 하고 있는 9명의 산모만이 존재

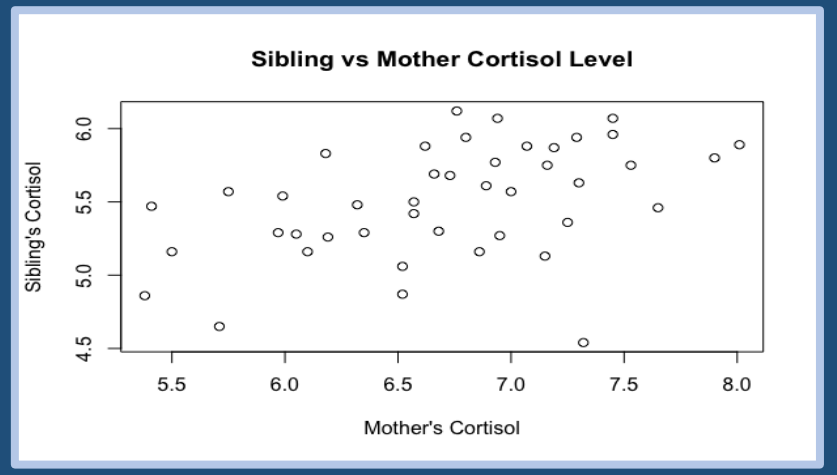
	Mother	Infant	Sibling
Sample Size	43	43	43
Sex (Males : Females)	0:43	16:27	27:16
Age (mean ± SE)	27.3 ± 0.73 (Years)	16.9 ± 1.16 (Days)	4.7 ± 0.4 (Years)
Cortisol Level (mean ± SE)	6.7 ± 0.1	4.2 ± 0.07	5.5 ± 0.06
Island (North : South: East)	8 : 23 : 12	8 : 23 : 12	8 : 23 : 12
Mother's Stress (mean ± SE)	3.2 ± 0.13	NA	NA
Mother Breastfeeding (Y : N)	9 : 34	NA	NA

분석기법 및 결과

- 상관관계 테스트, 가설검정(H_0 : 산모와 영아의 코티솔 레벨 사이에 관계가 없다 / H_a: 산모와 영아의 코티솔 레벨 사이에 관계가 있다)



- 그래프에 의하면 초기 가정과 다르게 산모와 영아보다 산모와 형제/자매 사이에서 강한 상관관계가 나타남



- 산모와 영아의 코티솔 수준과 산모와 형제간의 코티솔 수준의 상관관계가 매우 높지는 않지만, 산모의 코티솔 수준은 영아와 영아의 형제들의 코티솔 수준에 영향

감사합니다

포트폴리오에 작성된 분석의 원본은 개인 홈페이지에 업로드 했습니다.
https://github.com/BoramShim/PORTFOLIO