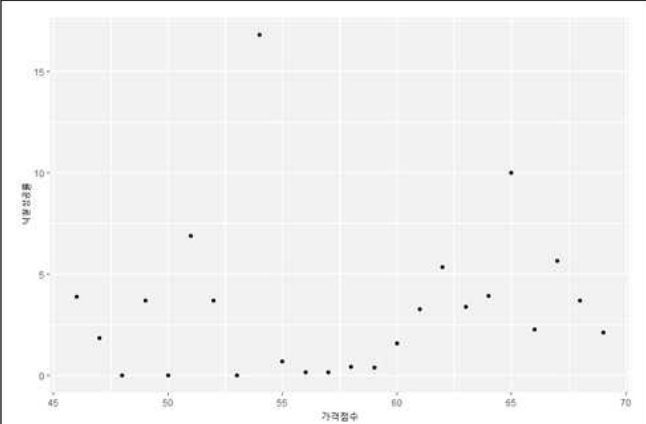
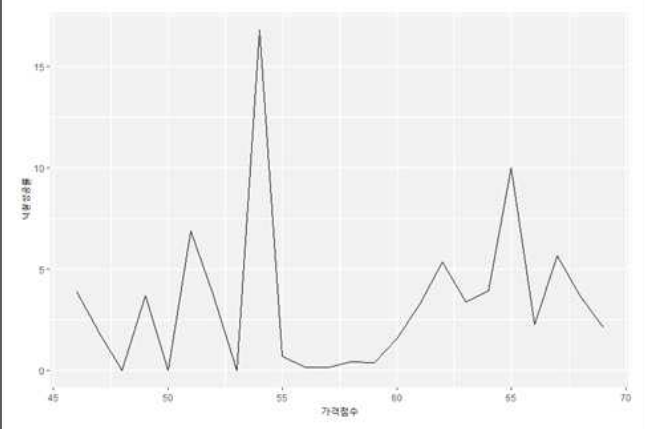
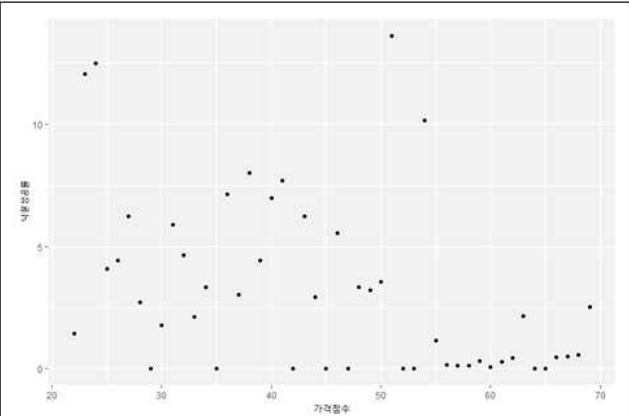
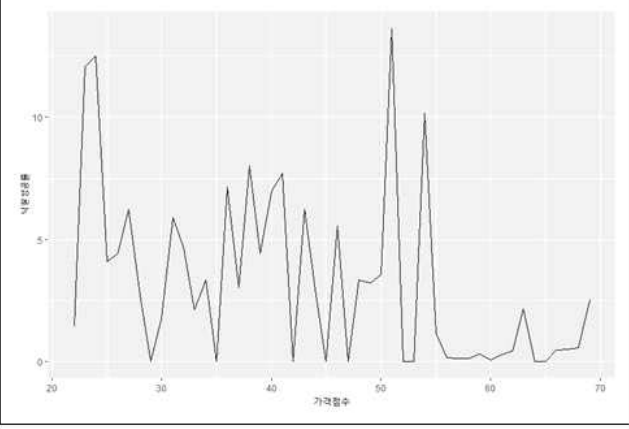


1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 22일								방법	Meeting				
활동	가격점수와 낙찰성공률의 상관관계 결과 정리								멤버	정진혁, 서지영				
[추정가격 10억원 미만]					[고시금액 미만]					[고시금액 미만]				
> show_result					> show_result_2					26	47	0	34	0.00000000
가격점수 낙찰성공수 낙찰참가수 낙찰성공률					가격점수 낙찰성공수 낙찰참가수 낙찰성공률					27	48	1	30	3.33333333
1	46	3	77	3.8961039	1	22	1	70	1.42857143	28	49	1	31	3.22580645
2	47	1	54	1.8518519	2	23	7	58	12.06896552	29	50	1	28	3.57142857
3	48	0	33	0.0000000	3	24	6	48	12.50000000	30	51	3	22	13.63636364
4	49	1	27	3.7037037	4	25	2	49	4.08163265	31	52	0	22	0.00000000
5	50	0	28	0.0000000	5	26	2	45	4.44444444	32	53	0	26	0.00000000
6	51	2	29	6.8965517	6	27	3	48	6.25000000	33	54	23	226	10.17699115
7	52	1	27	3.7037037	7	28	1	37	2.70270270	34	55	116	10074	1.15147905
8	53	0	16	0.0000000	8	29	0	54	0.00000000	35	56	15	8708	0.17225540
9	54	17	101	16.8316832	9	30	1	56	1.78571429	36	57	9	7057	0.12753295
10	55	27	3882	0.6955178	10	31	3	51	5.88235294	37	58	6	4509	0.13306720
11	56	5	2857	0.1750088	11	32	2	43	4.65116279	38	59	8	2450	0.32653061
12	57	3	1896	0.1582278	12	33	1	47	2.12765957	39	60	1	1283	0.07794232
13	58	3	704	0.4261364	13	34	1	30	3.33333333	40	61	2	714	0.28011204
14	59	1	259	0.3861004	14	35	0	48	0.00000000	41	62	2	448	0.44642857
15	60	2	128	1.5625000	15	36	2	28	7.14285714	42	63	6	280	2.14285714
16	61	2	61	3.2786885	16	37	1	33	3.03030303	43	64	0	309	0.00000000
17	62	3	56	5.3571429	17	38	2	25	8.00000000	44	65	0	323	0.00000000
18	63	2	59	3.3898305	18	39	2	45	4.44444444	45	66	2	421	0.47505938
19	64	2	51	3.9215686	19	40	3	43	6.97674419	46	67	3	587	0.51107325
20	65	5	50	10.0000000	20	41	2	26	7.69230769	47	68	5	911	0.54884742
21	66	1	44	2.2727273	21	42	0	22	0.00000000	48	69	32	1272	2.51572327
22	67	3	53	5.6603774	22	43	2	32	6.25000000					
23	68	2	54	3.7037037	23	44	1	34	2.94117647					
24	69	2	94	2.1276596	24	45	0	33	0.00000000					
25	70	0	0	NaN	25	46	2	36	5.55555556					

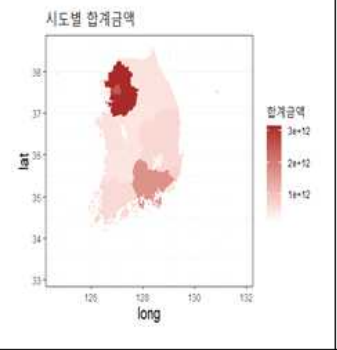
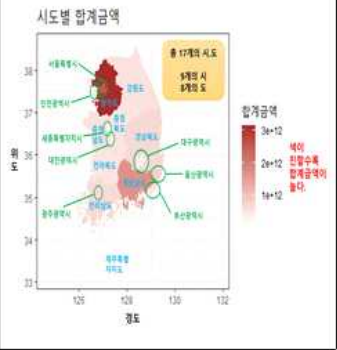
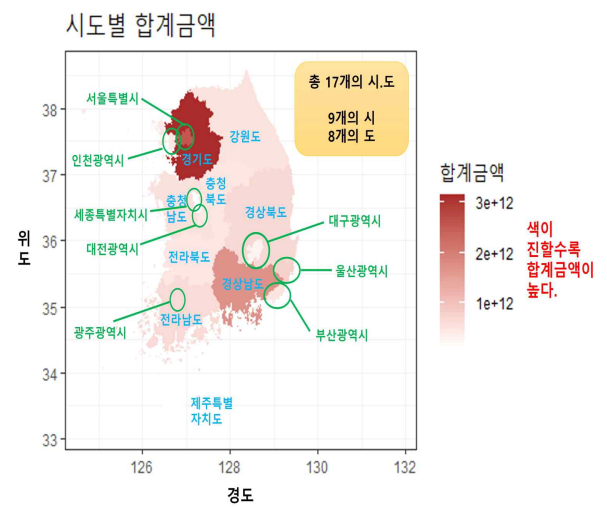
1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 22일	방법	Meeting
활동	가격점수와 낙찰성공률의 상관관계 결과 정리	멤버	정진혁, 서지영
<div> <div> <p>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</p>   </div> <div> <p>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</p>   </div> </div> <div> <p>[추정가격 10억원 미만]</p> <p>가격점수가 54점일 때와 65점일 때 가장 낙찰성공률이 높은 것으로 나타났다.</p> <p>낙찰성공률로만 판단하였을 때에는 55점이 가장 이상적이지만 분모인 낙찰참가수가 워낙 커서 낙찰성공률은 낮게 나타났다.</p> <p>가격요소와 비가격 요소로 평가가 이루어짐을 감안할 때 무조건 가격 점수가 높다고 하여 낙찰되는 것이 아님을 보여주는 결과라고 생각했다.</p> <p>[고시금액 미만]</p> <p>가격점수가 23점, 24점, 51점, 54점일 때 낙찰성공률이 높은 것으로 나타났다.</p> <p>낙찰성공률로만 판단하였을 때에는 55점이 가장 이상적이지만 분모인 낙찰참가수가 워낙 커서 낙찰성공률은 낮게 나타났다.</p> <p>비가격요소에서 평가를 잘 받을 계획을 세우고 가격 점수에서 55점 이상을 확보하는 것이 오히려 낙찰 확률을 높일 수도 있겠다는 생각을 했다.</p> </div>			
3-23. Rstudio 코드.hwp			

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 23일	방법	Zoom 회의
활동	앞으로 해야 할 일 계획 / 활동 보고	멤버	정진혁, 서지영
<p>[대주제 1 : 가격점수가 적격심사 낙찰에 미치는 영향]</p> <p>< 소주제 1 - 낙찰참가수와 가격점수와의 관계 ></p> <p>낙찰 참가수가 낮을 때에는 가격점수가 천차만별로 분포했지만 낙찰참가수가 증가할수록 점점 55점에 수렴하는 양상을 보였다.</p> <p>회귀분석을 실시하면 지수함수 또는 로그함수로 표현될 것이라 예상하고 있는데 이에 대한 추가 공부가 필요하다.</p> <p>< 소주제 2 - 가격점수와 낙찰성공률 ></p> <p>특정 가격점수에서 갑자기 낙찰성공률이 높은 극대점이 2군데 정도씩 있었다.</p> <p>추정가격이 10억원 미만일 때와 고시금액 미만일 때 모두 4차 함수 형태로 회귀분석을 실시하면 가격점수와 낙찰성공률의 관계를 구할 수 있을 것이라 예상하고 있다.</p> <p>이상적인 가격점수가 존재할 것이라 생각했으나 그래프를 보니 결론적으로는 가격점수 뿐만 아니라 비가격요소의 영향도 무시할 수 없다는 결론을 내렸다.</p>	<p>[대주제 2 : 조달 실적에서 공급 기업의 편중 현황]</p> <p>< 소주제 1 - 지역별 조달 실적의 편중 ></p> <p>< 소주제 2 - 기업구분별 조달 실적의 편중 ></p> <p>공급 기업의 편중이라는 주제를 다루기 위해서는 공급 기업을 어떤 기준으로 분류할 수 있는지 알아야 한다고 생각하였다.</p> <p>우선적으로 생각한 기준은 조달 기업이 위치한 지역과 기업의 종류(대기업, 중견기업, 중소기업, 기타 등)이다.</p> <p>조달 실적을 다룰 수 있는 데이터를 찾았는데 소주제 1에 활용할 수 있는 데이터를 찾았다.</p> <p>조달기업별 실적 순위라는 데이터였는데 업체의 소재지와 업체명이 나타난 데이터였다.</p> <p>건수와 금액에 대한 데이터가 나열되어 있었지만 건수보다는 금액이 더 중요한 데이터라는 생각을 했다.</p> <p>먼저 지역별로 데이터를 나누고 금액의 합을 구하여 어느 지역에 편중되어 있는지 지도로 시각화해보기로 계획을 세웠다.</p>	<p>[대주제 1 : 가격점수가 적격심사 낙찰에 미치는 영향]</p> <p>시각화한 자료들을 회귀분석을 실시해 그래프의 식을 구체적으로 구해야 하므로 회귀분석에 대한 공부가 추가적으로 필요하다.</p>	
		<p>[대주제 2 : 조달 실적에서 공급 기업의 편중 현황]</p> <p>- 소주제 1에 대한 지도 시각화 진행하기</p> <p>- 지역별로 금액 데이터 구분하고 지도에 그라데이션으로 편중 정도 나타내기</p> <p>- 소주제 2를 탐구할 수 있는 데이터 찾기</p> <p>- 소주제 1에 대해 지도 시각화 이외에 다르게 표현할 수 있는 방법이 있는지 생각해보기</p>	

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 24일	방법	Homework
활동	지역별 조달 실적의 편중 현황 - 지도 시각화	멤버	서지영
<p>1-33. 조달기업별 실적 순위 - Rstudio 코드</p> <pre># 금액 오름차순데이터(순위 오름차순)와 건수 오름차순 데이터 읽기 price_dat = read_xlsx("1-33. 조달기업별 실적 순위.xlsx") %>% data.frame(stringsAsFactors = F) price_dat = price_dat %>% slice(c(2:nrow(price_dat))) %>% filter(건수 > 0 & 금액 > 0) num_dat = price_dat %>% arrange(건수) price = price_dat\$금액 head(price, 10); tail(price, 10) # 지역별 합계 금액 데이터 unique(price_dat\$소재지) local_dat = price_dat %>% group_by(., 소재지) %>% summarise_at(vars(금액), sum, na.rm = T) %>% data.frame(stringsAsFactors = F) %>% rename(합계금액 = 금액) local_dat = local_dat %>% arrange(합계금액) local_dat = local_dat[c(1,4),] #미분류, 국외소재 제외 (17개 시도 구분으로 바꿈) # 지도 데이터 시각화 local_dat = local_dat[c(16,9,4,5,3,6,14,1,17,10,7,11,8,12,13,15,2),] rownames(local_dat) = c(1:nrow(local_dat)) local_dat\$id = 0:16 korea = shapefile("TL_SCCO_CTPRVN.shp") korea = spTransform(korea, CRS("+proj=longlat")) korea_map = fortify(korea) merge_result = merge(korea_map, local_dat, by = "id") plot = ggplot(data = merge_result) + geom_polygon(aes(x = long, y = lat, group = group, fill = 합계금액)) + labs(title = "시도별 합계금액") + theme_bw() + theme(text = element_text(size = 15)) plot = plot + scale_fill_gradient2(low = "lightyellow", mid = "white", high = "brown", midpoint = .02)</pre> <div>   </div>		<p>1-33. 데이터를 활용한 핸들링을 진행했다.</p> <p>건수와 금액이 음수인 것이 있어 이들에 대한 데이터는 제외하고 분석했다.(price_dat)</p> <p>먼저 unique한 소재지가 무엇인지 분석했고 총 9개의 시와 8개의 도로 이루어짐을 알았다.</p> <p>local_dat에 Excel 데이터를 소재지를 기준으로 금액에 대한 합계를 구한다.</p> <p>지도로 시각화를 구현하려면 소재지를 일정 순서로 정렬하는 것이 필요한데 이 때문에 local_dat의 행 순서를 모두 바꾸었다.</p> <div>  </div> <p>시도별 합계 금액을 구한 결과 서울특별시와 경기도, 경상남도가 합계 금액이 월등히 높은 것으로 나타났다.</p> <p>붉은 계열의 그라데이션을 사용하여 표현하였는데 색이 진할수록 합계 금액이 높다.</p> <p>다른 지역과 확연히 차이가 드러나며 지도에 각각 시도명을 표시했다.</p>	
https://coding-law.tistory.com/30		1-33. 조달기업별 실적 순위.xlsx	

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 24일	방법	Homework
활동	지역별 조달 실적의 편중 현황 - 편차 막대 그래프	멤버	서지영
<p>1-33. 조달기업별 실적 순위 - Rstudio 코드</p> <pre># 편차 데이터 시각화 local_dat = local_dat %>% mutate(편차 = 합계금액 - mean(합계금액)) sd_plot = ggplot(data = local_dat) + geom_bar(aes(x = 소재지, y = 편차, fill = as.factor(소재지)), stat = "identity") + theme(title = element_text(size = 15)) + theme_bw() + theme(axis.text.x = element_text(angle = 70, hjust = 1)) + labs(title = "시도별 합계금액 편차") + theme(legend.position = "none") sd_plot</pre>		<p>앞선 시각화에서는 지도를 통해 표현함으로써 어느 지역에 조달실적 금액이 얼마나 편중되었는지 나타냈다면 이번에는 다른 방면으로도 편중되어 있음을 보여주었다.</p> <p>편차 = 실제 데이터 - 평균 임을 이용하여 편차 그래프를 나타내고 y값인 편차가 0일 때 평균, 이 그래프에서 멀어질수록 편차가 심하다고 할 수 있다.</p> <p>그래프가 좌표축 0을 기준으로 위쪽에 위치하면 평균보다 크고 아래쪽에 위치하면 평균보다 작다.</p> <p>경기도, 경상남도, 서울특별시가 편차가 매우 큰 것으로 보아 압도적이었다.</p>	
		<p>지도 시각화를 통해 편차가 현재 심함을 알 수 있었다. 2가지 시각화 자료들을 통해 공급 기업의 편중 주제에서 지역을 기준으로 편중이 심하다는 것을 보여줄 수 있는 자료로 활용할 수 있을 것이라 생각한다.</p> <p>이는 다음 해에 수치적으로 어떨 것이라라고 예측하기보다는 현재의 상황을 반영하는 자료로 활용할 수 있을 것이다.</p>	
1-33. Rstudio 코드.hwp	시도별 합계금액 금액.jpeg	시도별 합계금액 편차.jpeg	

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 25일	방법	Homework
활동	가격점수와 낙찰성공률의 상관관계 - 회귀분석	멤버	서지영

회귀분석을 하기 위한 통계 모형의 4가지 가정

1. 선형성 평가(Residuals Vs Fitted)

종속 변수가 독립변수와 직선적 관계를 맺고 있다면, 잔차와 예측된 값 사이에는 체계적 관련성이 없어야 한다.

즉, 모델은 무작위 오차를 제외하고 데이터에 존재하는 모든 체계적 변량을 포착해야 한다.

2. 정규성 평가(Normal Q-Q)

정규성 가정에 충족한다면 정규성 그래프를 그렸을 때 점들이 45도 직선에 위치해야 한다. 직선 밖에서 멀리 떨어져 있다면 극단적인 데이터로 인해 정규성이 흔들릴 수 있음을 알 수 있다.

3. 등분산성 평가(Scale-Location)

등분산성을 만족하려면 수평선 주변에 무작위적으로 위치해야 한다.

4. 오차항들의 독립성(Residuals Vs Leverage)

개별 관측치(이상치, 큰 지레점, 영향치)에 대한 정보를 제공한다.

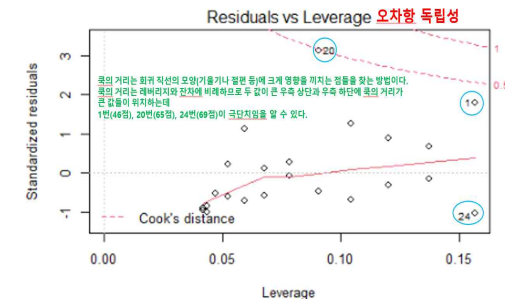
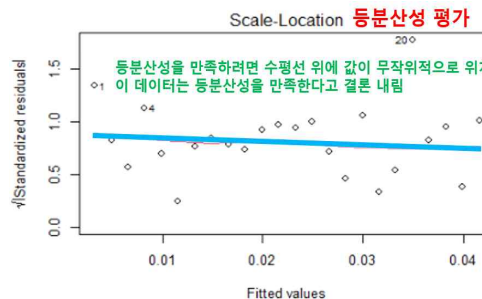
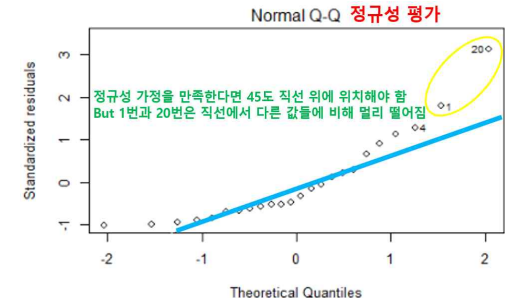
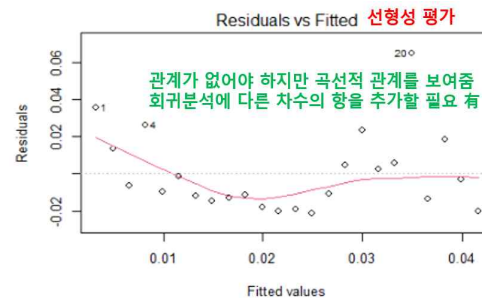
[고시금액 10억원 미만 - 단순선형회귀분석]

```
> res = lm(낙찰성공률~가격점수, data = final_dat)
> summary(res)

Call:
lm(formula = 낙찰성공률 ~ 가격점수, data = final_dat)

Residuals:
    Min       1Q   Median       3Q      Max
-0.021065 -0.014128 -0.008265  0.007866  0.065078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0733708  0.0371617  -1.974  0.0610
가격점수      0.0016660  0.0006417   2.596  0.0165 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
변동의 설명력(결정계수)          조정된 결정계수
Residual standard error: 0.02176 on 22 degrees of freedom
Multiple R-squared:  0.2346,    Adjusted R-squared:  0.1998
F-statistic: 6.742 on 1 and 22 DF, p-value: 0.01647
```



1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 25일	방법	Homework
활동	가격점수와 낙찰성공률의 상관관계 - 회귀분석	멤버	서지영
<div> <div> <p>[고시금액 10억원 미만 - 다항회귀분석]</p> <pre> > res = lm(낙찰성공률~가격점수 + I(가격점수^2) , data = final_dat[-20,]) > summary(res) Call: lm(formula = 낙찰성공률 ~ 가격점수 + I(가격점수^2), data = final_dat[-20,]) Residuals: Min 1Q Median 3Q Max -0.026420 -0.009318 -0.005587 0.011765 0.033303 Coefficients: (Intercept) 가격점수 I(가격점수^2) 5.016e-01 -1.826e-02 1.693e-04 P-value(제 1종의 오류를 범할 확률) (Intercept) 5.016e-01 2.350e-01 2.135 가격점수 -1.826e-02 8.257e-03 -2.211 I(가격점수^2) 1.693e-04 7.173e-05 2.360 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 변동의 설명력(결정계수) 조정된 결정계수 Residual standard error: 0.01501 on 20 degrees of freedom Multiple R-squared: 0.3854, Adjusted R-squared: 0.3239 F-statistic: 6.27 on 2 and 20 DF, p-value: 0.007694 </pre> </div> <div> </div> </div>			
<p>회귀분석을 통해 정리된 데이터를 점검하면서 이상치가 데이터 분포에 생각보다 많은 영향을 끼칠 수 있음을 알게 되었다.</p> <p>이상치가 있을 때 이를 어떻게 다루어야 할지 고민하는 시간이 필요할 것으로 보인다.</p> <p>지금까지 생각해본 해결책으로는 1. 이상치 삭제 / 2. 회귀분석의 차수 늘리기 / 3. 변수의 변환(더하거나 빼기 등) 이다.</p> <p>현재로서는 고시금액 10억원 미만에 대한 데이터에 대해 가격점수와 낙찰성공률에 대한 데이터로 진행했으나 낙찰참가수와 가격점수에 대한 데이터, 고시금액 미만 항목에 대하여도 비슷한 방식으로 진행하면 좋을 것 같았다.</p> <p>또한 가격점수 구간을 기존에는 1로 나누었었는데 이 구간을 좁혀서 즉, 데이터 개수를 늘려서 진행해보면 어떨까라는 생각을 하게 되었다.</p>			
회귀분석공부(20220125-서지영).docx			

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(22일 ~ 26일)

날짜	1월 26일	방법	Homework
활동	회귀진단 / 이상치 추출 방법 공부	멤버	서지영
< 회귀진단의 향상된 방법 >		< 이상치 추출 방법 - 흔치 않은 관측치 조정 방법 >	
<p>[1] 정규성 qqplot 함수 이용</p>	<p>[2] 오차 독립성</p> <ul style="list-style-type: none"> - 독립성 여부를 평가하기 위한 가장 좋은 방법이 자료가 어떻게 수집되었는가에서 출발한다. - Dubin - Waston 검정을 하여 유의미하지 않은 p값은 상관 정도가 낮으므로 오차들이 독립적이라고 말할 수 있다. 	<p>[1] 이상치 이상치는 모델로 잘 예측할 수 없는 관측치 대체로 양수나 음수의 큰 잔차를 가짐 outlierTest 함수를 이용하며 이상치의 p값이 유의미하다면 이상치를 삭제한 후 검정해야 한다.</p> <p>[2] 큰 지레점 다른 예측 변수 값들에 비추어 이상치인 경우</p> <p>[3] 영향치 n이 표본 크기, k가 예측변수인 경우에 $4/(n-k-1)$보다 Cook의 Distance가 크다면 영향치이다.</p> <p>* influencePlot함수를 이용하여 이상치, 지레점, 영향치 플롯에 대한 정보를 결합한 종합 정보를 시각화하여 보여준다.</p>	
	<p>[3] 선형성</p> <ul style="list-style-type: none"> - crplots 함수를 이용 - 플롯 결과 중 어느 하나라도 비선형성이 나타난다면 다항회귀와 같은 곡선적 요인을 추가, 변수 전환, 다른 회귀 분석 적용 등의 방법 모색이 필요하다. 		
회귀분석공부(20220126-서지영).docx			