

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 15일	방법	Homework																		
활동	통계적 분석 기법 - 회귀분석 공부	멤버	정진혁, 서지영																		
<div>< 회귀분석 ></div> <div>- 단순회귀 : 종속변수 y를 독립 변수 x의 함수로서 설명</div> <div>- 실험이나 관측에 따른 편차가 존재하며 정확히 수학적인 관계보다는 통계적인 관계가 성립(관측 오차가 발생함)</div> <div>- 두 변수의 관계를 연구하기 위해 우선 산점도를 그려본다.</div> <div>- 선형적인 관계 / 오차항들의 독립 / 등분산 / 정규분포라는 4가지 가정을 내포</div> <div>[단순 선형 회귀 모형]</div> <div>- y와 x의 관계가 직선인 경우 $y = \beta_0 + \beta_1 x$</div> <div>- 설명변수 x_i에 대응되는 Y_i는 평균이 $\beta_0 + \beta_1 x_i$이고 표준편차 σ를 가지는 정규 분포를 따른다.</div> <div>- 즉 $i = 1, \dots, n$에 대하여 $Y_i = \beta_0 + \beta_1 x_i + e_i$이고 e_i는 독립적으로 $N(0, \sigma^2)$을 따름</div> <div>- 주어진 데이터 $(x_i, y_i), i = 1, \dots, n$을 이용하여 미지의 계수 β_0, β_1를 추정</div> <div>[추정 원리 - 최소제곱법]</div> <div>아래 식에서 D를 최소화하는 모수값 $\hat{\beta}_0, \hat{\beta}_1$을 구한다. 이를 β_0, β_1의 최소제곱추정량이라 한다.</div> <div>$d_i = y_i - \beta_0 - \beta_1 x_i : \text{관측치와 예측치의 차이}$$D = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$</div>		<div>[기호 및 추정]</div> <table><tr><td>표본평균</td><td>$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i / \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$</td></tr><tr><td>x의 편차제곱합</td><td>$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$</td></tr><tr><td>y의 편차제곱합</td><td>$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$</td></tr><tr><td>xy의 편차제곱합</td><td>$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$</td></tr><tr><td>기울기 및 y절편</td><td>$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} / \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$</td></tr><tr><td>추정된 회귀직선</td><td>$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$</td></tr><tr><td>잔차</td><td>$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i (\text{단}, i = 1, \dots, n)$</td></tr><tr><td>잔차(오차)제곱합</td><td>$SSE = \sum_{i=1}^n \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$</td></tr><tr><td>분산($\sigma^2$)의 추정량</td><td>$S^2 = \frac{SSE}{n-2}$</td></tr></table> <div>[직선 모형의 적합도 - 선형관계의 정도]</div> <div>$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} (\text{단}, r \text{은 표본 상관계수})$</div>		표본평균	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i / \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$	x의 편차제곱합	$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$	y의 편차제곱합	$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$	xy의 편차제곱합	$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$	기울기 및 y절편	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} / \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	추정된 회귀직선	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$	잔차	$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i (\text{단}, i = 1, \dots, n)$	잔차(오차)제곱합	$SSE = \sum_{i=1}^n \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$	분산(σ^2)의 추정량	$S^2 = \frac{SSE}{n-2}$
표본평균	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i / \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$																				
x의 편차제곱합	$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$																				
y의 편차제곱합	$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$																				
xy의 편차제곱합	$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$																				
기울기 및 y절편	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} / \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$																				
추정된 회귀직선	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$																				
잔차	$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i (\text{단}, i = 1, \dots, n)$																				
잔차(오차)제곱합	$SSE = \sum_{i=1}^n \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$																				
분산(σ^2)의 추정량	$S^2 = \frac{SSE}{n-2}$																				
R을 활용한 통계학 이론과 응용(제4판) - 자유아카데미 - 김동희 외 8인 - 10장 회귀분석 참고																					

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 16일	방법	Zoom 회의
활동	3-23 데이터 핸들링(추정가격 10억 이상) - 가격점수, 부적격 사유 분석	멤버	정진혁, 서지영
<p>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</p> <pre> # 필요한 패키지 불러오기 rm(list = ls()); gc(reset = T) require(readxl); require(dplyr) # 엑셀 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부) Excel = read_xlsx("/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱O).xlsx") %>% data.frame(stringsAsFactors = F) # 추정가격 10억원 이상인 데이터(139개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱 criteria = grep("10억원이상", Excel[, 1]); sub_dat = Excel[criteria, c(1, 3, 5, 6, 7, 8)] # 항목이 몇 개인지 확인(unique_item : 낙찰공고(13개), pick : 낙찰(4개)) unique_item = unique(sub_dat[, 2]); pick = grep("Y", sub_dat[, 6]) length(unique_item); length(pick) # 부적격사유 분석(fail_dat : 28행 3열[입찰률, 부적격여부, 부적격사유] / 입찰률이 NA인 행에 0 삽입) fail_item = grep("Y", sub_dat[, 4]) fail_dat = sub_dat[fail_item, c(3, 4, 5)] nrow(fail_dat) for(i in 1:nrow(fail_dat)){ if(is.na(fail_dat[i, 1]) == TRUE){ fail_dat[i, 1] = 0 }else{ fail_dat[i, 1] = fail_dat[i, 1] } } # 낙찰하한선 미달 개수(22개) fail_dat = fail_dat[grep("낙찰", fail_dat[, 3]), c(1, 3)] nrow(fail_dat); summary(fail_dat[, 1]) # 부적격이 아닌 데이터(111개)들의 입찰률(입찰가격/예정가격) good_item = sub_dat[grep("N", sub_dat[, 4]), c(1, 3)] nrow(good_item); summary(good_item) # 부적격이 아닌 데이터(111개)들의 가격점수 구하기 price_score = sub_dat[grep("N", sub_dat[, 4]), c(3, 6)] %>% data.frame(stringsAsFactors = F) sc = price_score %>% mutate(score = c(1:nrow(price_score))) for(i in 1:nrow(sc)){ sc[i, 3] = 55-2*abs(88-sc[i, 1]) } summary(sc[, 3]) # 부적격이 아닌 데이터들 중 낙찰된 데이터(4개)들의 가격점수 구하기 sub_sc = sc[grep("Y", sc[, 2]), c(1, 3)] summary(sub_sc) </pre>		<p>[코딩 알고리즘 순서]</p> <ol style="list-style-type: none"> 엑셀 데이터 읽기 <ul style="list-style-type: none"> 총 8열로 낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부로 이루어진 데이터이다. 낙찰공고의 개수, 낙찰된 개수 구하기 <ul style="list-style-type: none"> 같은 공고의 경우 예정가격이 같다는 원리를 이용해 unique 기능을 이용하여 낙찰공고의 개수를 구하였다. 8열의 낙찰여부는 N(낙찰 X), Y(낙찰 O)로 이루어져 있으므로 Y인 행의 개수를 찾으면 낙찰된 데이터 개수를 알 수 있다. 부적격 사유 분석하기 <ul style="list-style-type: none"> 부적격 사유는 낙찰여부와 마찬가지로 N과 Y로 이루어져 있고 Y인 행을 찾는다. 부적격 사유에서 "낙찰하한선 미달"인 데이터를 grep 기능으로 인덱싱하여 찾는다. 낙찰하한선 미달인 데이터들의 입찰률(입찰가격/예정가격) summary 부적격이 아닌 나머지 데이터들의 가격점수 구하기 <ul style="list-style-type: none"> 입찰률 열과 가격점수를 구하는 수식을 이용해 가격점수를 구하는 열을 mutate 기능을 이용해 삽입한다. 낙찰된 데이터들의 가격점수와 부적격 판정을 받지 않았지만 낙찰되지 않은 데이터들의 가격점수를 summary를 이용해 비교해본다. 	
3-23. 물품 입찰분류별 투찰업체 내역(2021_12~2022_01).xlsx		3-23. 물품입찰분류별 투찰업체 내역(인덱싱O).xlsx	

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 17일	방법	Homework
활동	3-23 데이터 핸들링(10억원 미만, 추정가격 미만) - 가격점수, 부적격 사유 분석	멤버	정진혁

3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드

```
# 필요한 패키지 불러오기
rm(list = ls()); gc(reset = T)
require(readxl); require(dplyr)

# 엑셀 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부)
Excel = read_xlsx("/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱0).xlsx") %>% data.frame(stringsAsFactors = F)

# 추정가격 10억원 미만인 데이터(19527개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱
criteria = grep("10억원미만", Excel[, 1]); criteria_2 = grep("10억원 미만", Excel[, 1])
sub_dat = Excel[c(criteria, criteria_2), c(1, 3, 5, 6, 7, 8)]
nrow(sub_dat)

# 항목이 몇 개인지 확인(unique_item : 낙찰공고(200개), pick : 낙찰(88개))
unique_item = unique(sub_dat[, 2]); pick = grep("Y", sub_dat[, 6])
length(unique_item); length(pick)

# 부적격사유 분석(fail_dat : 8887행 3열[입찰률, 부적격여부, 부적격사유] / 입찰률이 NA인 행에 0 삽입)
fail_item = grep("Y", sub_dat[, 4])
fail_dat = sub_dat[fail_item, c(3, 4, 5)]
nrow(fail_dat)

for(i in 1:nrow(fail_dat)){
  if(is.na(fail_dat[i, 1]) == TRUE){
    fail_dat[i, 1] = 0
  }else{
    fail_dat[i, 1] = fail_dat[i, 1]
  }
}

# 낙찰제한선 미달 개수(8440개)
fail_dat = fail_dat[grep("낙찰", fail_dat[, 3]), c(1, 3)]
nrow(fail_dat); summary(fail_dat[, 1])

# 부적격이 아닌 데이터(10640개)들의 입찰률(입찰가격/예정가격)
good_item = sub_dat[grep("N", sub_dat[, 4]), 3]
length(good_item); summary(good_item)

# 부적격이 아닌 데이터(10640개)들의 가격점수 구하기
price_score = sub_dat[grep("N", sub_dat[, 4]), c(3, 6)] %>% data.frame(stringsAsFactors = F)
sc = price_score %>% mutate(score = c(1:nrow(price_score)))
for(i in 1:nrow(sc)){
  sc[i, 3] = 55-2*abs(88-sc[i, 1])
}
summary(sc[, 3])

# 부적격이 아닌 데이터들 중 낙찰된 데이터(88개)들의 가격점수 구하기
sub_sc = sc[grep("Y", sc[, 2]), c(1, 3)]
summary(sub_sc)
```

3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드

```
# 필요한 패키지 불러오기
rm(list = ls()); gc(reset = T)
require(readxl); require(dplyr)

# 엑셀 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부)
Excel = read_xlsx("/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱0).xlsx") %>% data.frame(stringsAsFactors = F)

# 추정가격 고시금액미만인 데이터(76683개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱
criteria = grep("고시금액미만", Excel[, 1])
sub_dat = Excel[criteria, c(1, 3, 5, 6, 7, 8)]
nrow(sub_dat)

# 항목이 몇 개인지 확인(unique_item : 낙찰공고(593개), pick : 낙찰(283개))
unique_item = unique(sub_dat[, 2]); pick = grep("Y", sub_dat[, 6])
length(unique_item); length(pick)

# 부적격사유 분석(fail_dat : 35876행 3열[입찰률, 부적격여부, 부적격사유] / 입찰률이 NA인 행에 0 삽입)
fail_item = grep("Y", sub_dat[, 4])
fail_dat = sub_dat[fail_item, c(3, 4, 5)]
nrow(fail_dat)

for(i in 1:nrow(fail_dat)){
  if(is.na(fail_dat[i, 1]) == TRUE){
    fail_dat[i, 1] = 0
  }else{
    fail_dat[i, 1] = fail_dat[i, 1]
  }
}

# 낙찰제한선 미달 개수(33516개)
fail_dat = fail_dat[grep("낙찰", fail_dat[, 3]), c(1, 3)]
nrow(fail_dat); summary(fail_dat[, 1])

# 부적격이 아닌 데이터(40806개)들의 입찰률(입찰가격/예정가격)
good_item = sub_dat[grep("N", sub_dat[, 4]), 3]
length(good_item); summary(good_item)

# 부적격이 아닌 데이터(40806개)들의 가격점수 구하기
price_score = sub_dat[grep("N", sub_dat[, 4]), c(3, 6)] %>% data.frame(stringsAsFactors = F)
sc = price_score %>% mutate(score = c(1:nrow(price_score)))
for(i in 1:nrow(sc)){
  sc[i, 3] = 55-2*abs(88-sc[i, 1])
}
summary(sc[, 3])

# 부적격이 아닌 데이터들 중 낙찰된 데이터(283개)들의 가격점수 구하기
sub_sc = sc[grep("Y", sc[, 2]), c(1, 3)]
summary(sub_sc)
```

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 18일	방법	Homework																								
활동	데이터 핸들링한 결과 정리하기	멤버	정진혁																								
<table> <tr> <th></th><th>추정가격이 10억원 이상</th><th>추정가격이 고시금액 이상 10억원 미만</th><th>추정가격이 고시금액 미만</th></tr> <tr> <td>낙찰공고 개수</td><td>13</td><td>200</td><td>593</td></tr> <tr> <td>낙찰 참가자 수</td><td>139</td><td>19527</td><td>76683</td></tr> <tr> <td>실제 낙찰된 데이터 개수</td><td>4</td><td>88</td><td>283</td></tr> <tr> <td>부적격판정 데이터 개수</td><td>28</td><td>8887</td><td>35876</td></tr> <tr> <td>낙찰하한선 미달 데이터 개수</td><td>22</td><td>8440</td><td>33516</td></tr> </table>					추정가격이 10억원 이상	추정가격이 고시금액 이상 10억원 미만	추정가격이 고시금액 미만	낙찰공고 개수	13	200	593	낙찰 참가자 수	139	19527	76683	실제 낙찰된 데이터 개수	4	88	283	부적격판정 데이터 개수	28	8887	35876	낙찰하한선 미달 데이터 개수	22	8440	33516
	추정가격이 10억원 이상	추정가격이 고시금액 이상 10억원 미만	추정가격이 고시금액 미만																								
낙찰공고 개수	13	200	593																								
낙찰 참가자 수	139	19527	76683																								
실제 낙찰된 데이터 개수	4	88	283																								
부적격판정 데이터 개수	28	8887	35876																								
낙찰하한선 미달 데이터 개수	22	8440	33516																								
추정가격이 10억원 이상		추정가격이 고시금액 이상 10억원 미만	추정가격이 고시금액 미만																								
<pre>> summary(fail_dat[, 1]) Min. 1st Qu. Median Mean 3rd Qu. Max. 79.72 80.22 80.37 80.28 80.42 80.46 > summary(good_item) Min. 1st Qu. Median Mean 3rd Qu. Max. 80.50 80.88 81.27 84.28 84.60 99.47 > summary(sc[, 3]) Min. 1st Qu. Median Mean 3rd Qu. Max. 32.07 40.49 41.16 42.09 42.36 54.95 > summary(sub_sc) 입찰률 score Min. :80.53 Min. :40.06 1st Qu.:81.16 1st Qu.:41.33 Median :82.39 Median :42.86 Mean :84.71 Mean :42.90 3rd Qu.:85.94 3rd Qu.:44.43 Max. :93.52 Max. :45.82</pre>		<pre>> summary(fail_dat[, 1]) Min. 1st Qu. Median Mean 3rd Qu. Max. 0.001 79.746 80.038 79.651 80.272 80.494 > summary(good_item) Min. 1st Qu. Median Mean 3rd Qu. Max. 80.50 80.81 81.22 82.21 81.78 100.00 > summary(sc[, 3]) Min. 1st Qu. Median Mean 3rd Qu. Max. 46.00 55.52 56.34 56.76 57.38 70.00 > summary(sub_sc) 입찰률 score Min. :80.50 Min. :46.00 1st Qu.: 80.52 1st Qu.:55.00 Median : 81.17 Median :55.22 Mean : 85.07 Mean :57.54 3rd Qu.: 89.79 3rd Qu.:61.02 Max. :100.00 Max. :69.88</pre>	<pre>> summary(fail_dat[, 1]) Min. 1st Qu. Median Mean 3rd Qu. Max. 0.00 83.52 83.88 77.58 84.12 89.97 > summary(good_item) Min. 1st Qu. Median Mean 3rd Qu. Max. 84.25 84.50 84.80 85.55 85.27 100.00 > summary(sc[, 3]) Min. 1st Qu. Median Mean 3rd Qu. Max. 22.00 55.86 57.02 57.56 58.67 70.00 > summary(sub_sc) 입찰률 score Min. :84.25 Min. :22.80 1st Qu.:84.27 1st Qu.:54.99 Median :84.52 Median :55.18 Mean :87.56 Mean :53.95 3rd Qu.:88.32 3rd Qu.:57.54 Max. :99.80 Max. :70.00</pre>																								
1) 낙찰하한선 미달 데이터들의 입찰률 / 2) 부적격 판정을 받지 않은 데이터들의 입찰률 / 3) 부적격 판정을 받지 않은 데이터들의 가격점수 / 4) 낙찰된 데이터들의 가격점수																											
3-23. Rstudio 코드.hwp																											

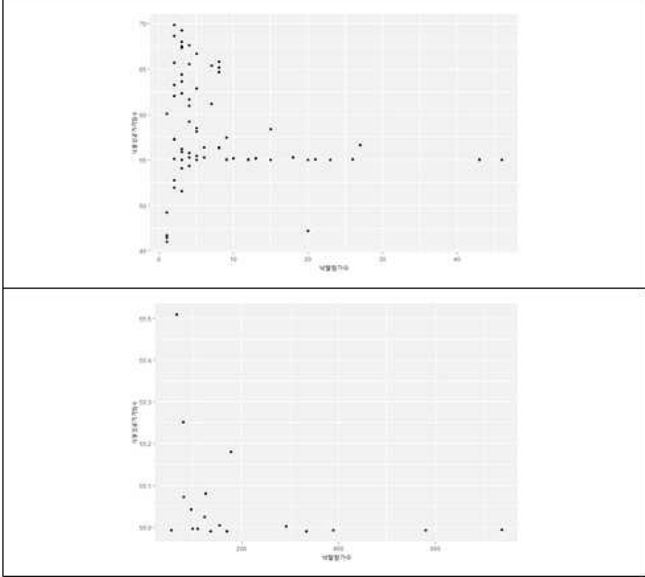
1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 18일	방법	Homework
활동	데이터 핸들링한 결과 정리하기	멤버	정진혁
<p>[1] 추정가격이 10억원 이상 / [2] 추정가격이 고시금액 이상 10억원 미만 / [3] 추정가격이 고시금액 미만</p> <p>[1]과 [2]의 경우 낙찰하한률이 80.50 정도라는 것을 알 수 있었다.</p> <p>낙찰하한선 미달 데이터들을 살펴보면 입찰률의 최댓값이 각각 80.46, 80.494였으며 부적격판정을 받지 않은 데이터 입찰률의 최솟값은 80.50이었다.</p> <p>[1]의 경우 부적격판정을 받지 않은 데이터들의 가격 점수는 평균은 낙찰된 데이터들과 비슷하지만 54.95점을 받은 데이터가 있었다.</p> <p>그러나 실제 낙찰된 데이터들을 보면 40점에서 45점 사이의 가격 점수를 받아도 낙찰이 되는 것을 알 수 있다.</p> <p>한편 [2]와 [3]으로 갈수록 가격점수의 편차가 심해짐을 알 수 있었다.</p> <p>가격 점수의 배점이 70점으로 [1]과 달리 15점 높아질 뿐만 아니라 입찰가격을 잘못 내놓을 경우 가격 점수에서의 감점이 큼을 분석할 수 있다.</p> <p>데이터를 분석하기 전에는 60점 이상을 받아야 안정적으로 낙찰이 될 수 있을 것이라 생각했지만 실제로는 그렇지 않았다.</p> <p>가격을 너무 낮게 잡아서 부적격 판정을 받는 것보다는 가격을 안정적으로 잡는 것이 낙찰 성공률을 높이는데 유리하다는 것을 알 수 있었다.</p>			
3-23. Rstudio 코드.hwp			

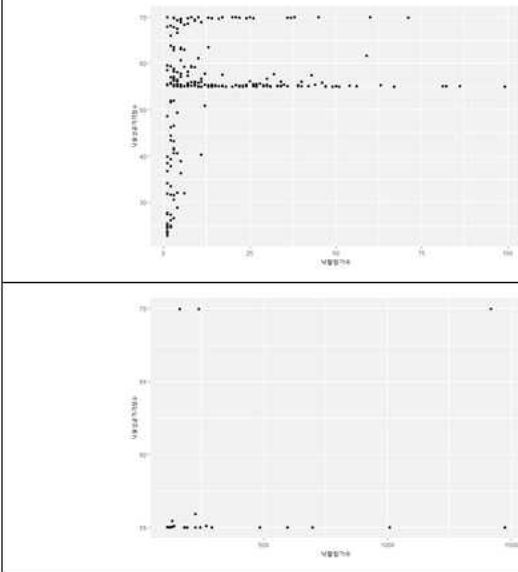
1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 19일	방법	Homework
활동	통계적 분석 기법 – 로지스틱 회귀	멤버	정진혁
<p>[로지스틱 회귀]</p> <p>이벤트가 발생할 확률을 결정하는 데 사용되는 통계 모델이다.</p> <p>특성 간의 관계를 보여주고 특정 결과의 확률을 계산한다.</p> <p>대상 변수가 이진이며 값은 1 또는 0이다.</p> <p>측정 가능한 항목에는 설명 변수/특성(측정 대상 항목)과 결과인 응답 변수/목표 이진 변수의 두 가지 유형이 있습니다.</p> <p>[로지스틱 회귀 종류]</p> <ul style="list-style-type: none">- 이진 로지스틱 회귀 : 범주형 응답에 대해 가능한 두 가지 결과- 다항 로지스틱 회귀 : 응답 변수에 순서가 없는 3개 이상의 변수가 포함- 순서 로지스틱 회귀 : 다항 회귀와 마찬가지로 3개 이상의 변수가 있으나 순서 O		<p>[로지스틱 회귀에 사용되는 가정]</p> <ul style="list-style-type: none">- 이진 로지스틱 회귀에서는 응답 변수가 이진이어야 합니다. 결과는 둘 중 어느 하나입니다.- 원하는 결과는 응답 변수의 요인 수준 1로 표시되어야 하며 원하지 않는 결과는 0입니다.- 의미를 가지는 변수만 포함해야 합니다.- 독립변수는 본질적으로 서로 독립적이어야 합니다. 다중 공선성이 거의 또는 전혀 없어야 합니다.- 로그 오즈와 독립 변수는 선형적으로 관련되어야 합니다.- 로지스틱 회귀는 대규모 크기 샘플에만 적용해야 합니다. <p>[로지스틱 회귀의 적용 분야]</p> <p>보건 의료, 정치, 제품 테스트, 마케팅, 금융 부문, 전자상거래 등</p>	
<p>로지스틱 회귀 분석을 알기 전까지는 선형 회귀 분석 기법을 이용하여 가격 점수와 낙찰 성공률에 대해 분석하려고 하였다.</p> <p>가격 점수를 구간으로 나누고 그에 따른 데이터 개수를 핸들링 하는 것 이외에 다른 방안을 찾아보다가 로지스틱 회귀 분석 기법을 알게 되었다.</p> <p>낙찰 성공을 1, 낙찰 실패를 0이라고 하고 가격 점수를 독립변수라고 하여 로지스틱 회귀 분석 기법을 실시하면 상관관계가 어느 정도인지 파악할 수 있을 것이란 생각이 들었다.</p> <p>앞선 데이터 핸들링에서 고시가격 10억원 이상에 대한 데이터는 수가 적어서 10억원 미만과 고시금액 미만에 대해 실시하기로 했다.</p>			
https://www.tibco.com/ko/reference-center/what-is-logistic-regression			

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 20일	방법	Meeting
활동	낙찰 참가수와 가격점수의 상관관계 코딩	멤버	정진혁, 서지영
<p>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</p> <pre> # 필요한 패키지 불러오기 rm(list = ls()); gc(reset = T) require(readr); require(dplyr) # 엑셀 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰율, 부적격여부, 부적격사유, 낙찰여부) Excel = read_excel("~/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱O).xlsx") %>% data.frame(stringsAsFactors = F) # 추정가격 10억원 미만인 데이터(19527개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱 criteria = grep("10억원미만", Excel[, 1]); criteria_2 = grep("10억원 미만", Excel[, 1]) sub_dat = Excel[c(criteria, criteria_2), c(1, 3, 5, 6, 7, 8)] nrow(sub_dat) # sub_dat : 예정가격 순으로 오름차순 정렬, 가격점수도 있음 sub_dat = sub_dat %>% arrange(sub_dat[, 2]) sub_dat = sub_dat %>% filter(부적격여부 == "N") sub_dat = sub_dat[, c(2, 3, 6)] sub_dat = sub_dat %>% mutate(score = c(1:nrow(sub_dat))) for(i in 1:nrow(sub_dat)){ sub_dat[i, 4] = 70-2*abs(8-sub_dat[i, 2]) } # 예정가격 unique 값 : unique_mom unique_mom = unique(sub_dat[, 1]) # mom_list에 예정가격 기준으로 데이터 분류 mom_list = list() for(i in 1:length(unique_mom)){ mom_list[[i]] = sub_dat %>% filter(예정가격 == unique_mom[i]) } # mom_length : 예정가격별 낙찰자 수 / pick_score : 낙찰 성공한 가격점수 mom_length = c(); pick_score = list() for(i in 1:length(mom_list)){ mom_length[i] = nrow(mom_list[[i]]) pick_score[i] = mom_list[[i]] %>% filter(낙찰여부 == "Y") pick_score[i] = pick_score[i][, 4] } for(i in 1:length(pick_score)){ if(length(pick_score[i]) == 0){ pick_score[i] = 0 }else{ pick_score[i] = pick_score[i][0] } } pick_score = pick_score %>% do.call(rbind, .) </pre>		<p>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</p> <pre> # fianl_dat : 낙찰참가수와 낙찰성공가격점수(낙찰 성공한 데이터들만 Pick) final_dat = data.frame(낙찰참가수 = mom_length, 낙찰성공가격점수 = pick_score) final_dat = final_dat %>% filter(낙찰성공가격점수 > 0) final_dat = final_dat %>% arrange(낙찰참가수) final_dat = final_dat %>% apply(,2, as.numeric) %>% data.frame(stringsAsFactors = F) final_dat_1 = final_dat %>% filter(낙찰참가수 < 50) final_dat_2 = final_dat %>% filter(낙찰참가수 > 50) # 낙찰참가수와 낙찰성공가격점수 간의 관계(ggplot2 시각화) require(ggplot2) ggplot(data = final_dat_1) + geom_point(aes(x = 낙찰참가수, y= 낙찰성공가격점수)) ggplot(data = final_dat_2) + geom_point(aes(x = 낙찰참가수, y= 낙찰성공가격점수)) </pre> 	
		<p>추정가격이 10억원 미만인 데이터에 대해 낙찰 참가수와 가격점수와의 상관관계를 분석하였다.</p> <p>먼저 부적격 판정을 받지 않은 데이터들을 뽑았다.</p> <p>같은 조달 공고이면 예정가격이 같다는 성질을 이용하여 unique_mom에 예정가격을 저장했다.</p> <p>mom_list에는 고유한 예정가격에 해당하는 데이터들만 순서대로 차례로 인덱싱하였다.</p> <p>mom_list에 저장된 각각의 원소들에 대해 length 함수를 적용하면 한 입찰공고에 대해 참가한 수를 알 수 있다.</p> <p>pick_score에는 낙찰에 성공한 가격점수 데이터를 삽입했다.</p> <p>fianl_dat는 mom_length와 pick_score로 이루어진 데이터 프레임이며 낙찰참가수의 편차가 커서 fianl_dat_1과 fianl_dat_2로 나누어 시각화를 구현했다.</p> <p>지수함수 또는 로그함수로 표현될 것으로 예상되며 55 점에 점차 수렴하는 양상을 보였다.</p>	

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 20일	방법	Meeting
활동	낙찰 참가수와 가격점수의 상관관계 코딩	멤버	정진혁, 서지영
<div>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</div> <div><pre># 필요한 패키지 불러오기 rm(list = ls()); gc(reset = T) require(readxl); require(dplyr) # 엑셀 문서 읽기(낙찰자점정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부) Excel = read_xlsx("/조달청/3-23. 물품입찰분류별 투찰업체 내역(민역상O).xlsx") %>% data.frame(stringsAsFactors = F) # 고시금액미만인 데이터(40806개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱 criteria = grep("고시금액미만", Excel[, 1]) sub_dat = Excel[criteria, c(1, 3, 5, 6, 7, 8)] nrow(sub_dat) # sub_dat : 예정가격 순으로 오름차순 정렬, 가격점수도 있음 sub_dat = sub_dat %>% arrange(sub_dat[, 2]) sub_dat = sub_dat %>% filter(부적격여부 == "N") sub_dat = sub_dat[, c(2, 3, 6)] sub_dat = sub_dat %>% mutate(score = c(1:nrow(sub_dat))) for(i in 1:nrow(sub_dat)){ sub_dat[i, 4] = 70-4*abs(8-sub_dat[i, 2]) } # 예정가격 unique 값 : unique_mom unique_mom = unique(sub_dat[, 1]) # mom_list에 예정가격 기준으로 데이터 분류 mom_list = list() for(i in 1:length(unique_mom)){ mom_list[[i]] = sub_dat %>% filter(예정가격 == unique_mom[i]) } # mom_length : 예정가격별 낙찰자 수 / pick_score : 낙찰 성공한 가격점수 mom_length = c(); pick_score = list() for(i in 1:length(mom_list)){ mom_length[i] = nrow(mom_list[[i]]) pick_score[i] = mom_list[[i]] %>% filter(낙찰여부 == "Y") pick_score[i] = pick_score[i][, 4] } for(i in 1:length(pick_score)){ if(length(pick_score[[i]]) == 0){ pick_score[[i]] = 0 }else{ pick_score[[i]] = pick_score[[i]] } } pick_score = pick_score %>% do.call(rbind, .)</pre></div>		<div>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</div> <div><pre># final_dat : 낙찰참가수와 낙찰성공가격점수(낙찰 성공한 데이터들만 Pick) final_dat = data.frame(낙찰참가수 = mom_length, 낙찰성공가격점수 = pick_score) final_dat = final_dat %>% filter(낙찰성공가격점수 > 0) final_dat = final_dat %>% arrange(낙찰참가수) final_dat = final_dat %>% apply(,2, as.numeric) %>% data.frame(stringsAsFactors = F) final_dat_1 = final_dat %>% filter(낙찰참가수 < 100) final_dat_2 = final_dat %>% filter(낙찰참가수 > 100) # 낙찰참가수와 낙찰성공가격점수 간의 관계(ggplot2 시각화) require(ggplot2) ggplot(data = final_dat_1) + geom_point(aes(x = 낙찰참가수, y = 낙찰성공가격점수)) ggplot(data = final_dat_2) + geom_point(aes(x = 낙찰참가수, y = 낙찰성공가격점수))</pre></div> <div></div>	
<p>고시금액 미만인 데이터들에 대해서도 같은 방법으로 데이터 분석을 진행하였다.</p> <p>낙찰참가수가 적을 때는 낙찰되는 가격점수가 1자 형태로 다양하게 분포되어 있던 반면 낙찰 참가수가 증가할수록 점차 55점에 수렴하는 양상을 보였다.</p> <p>추정가격이 10억원 미만일 때와 마찬가지로 지수함수 또는 로그함수로 표현될 것이라 예상하고 있다.</p> <p>낙찰 참가수가 다양하게 분포되어 있었는데 구분 기준을 100으로 하여 final_dat_1과 final_dat_2로 나누어 시각화를 진행하였다.</p> <p>추정가격이 10억원 미만일 때에는 낙찰 참가수의 구분 기준을 50으로 하였다.</p>			
3-23. Rstudio 코드.hwp			

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 21일	방법	Meeting
활동	가격점수와 낙찰성공률의 상관관계 코딩	멤버	정진혁, 서지영

3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드

```
# 필요한 패키지 불러오기
rm(list = ls()); gc(reset = T)
require(readxl); require(dplyr)

# 역별 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부)
Excel = read_xlsx("~/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱0).xlsx") %>% data.frame(stringsAsFactors = F)

# 추정가격 10억원 미만인 데이터(19527개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱
criteria = grep("10억원미만", Excel[, 1]); criteria_2 = grep("10억원 미만", Excel[, 1])
sub_dat = Excel[c(criteria, criteria_2), c(1, 3, 5, 6, 7, 8)]

price_dat = sub_dat[, c(3, 4, 6)]
price_dat = price_dat %>% mutate(score = c(1:nrow(price_dat)))
for(i in 1:nrow(price_dat)){
  price_dat[i, 4] = 70-2*abs(88-price_dat[i, 1])
}

# score_decision : 낙찰된 것 / 안된 것 - 가격점수로 오름차순 정렬
score_decision = price_dat[, c(2, 3, 4)]
score_decision = score_decision %>% arrange(score)
score_decision = score_decision %>% filter(부적격여부 == "N")

# 낙찰된 데이터만 가격점수 구간별로 나눔(my_list)
score_decision_Yes = score_decision %>% filter(낙찰여부 == "Y")
score_decision_Yes

my_list = list()
for(i in 1:25){
  group = list()
  for(j in 1:nrow(score_decision_Yes)){
    if(score_decision_Yes[j, 3] >= i+45 & score_decision_Yes[j, 3] < i+46){
      group[j] = score_decision_Yes[j, ]
    }else{
      group[j] = data.frame(부적격여부 = "N", 낙찰여부 = "Y", score = 0)
    }
  }
  group = group %>% do.call(rbind, .)
  my_list[[i]] = group %>% filter(score > 0)
}
my_list
```

3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드

```
# 낙찰된 것 / 안된 것 모두 가격점수 구간별로 나눔(my_list_2)
my_list_2 = list()
for(i in 1:25){
  group = list()
  for(j in 1:nrow(score_decision)){
    if(score_decision[j, 3] >= i+45 & score_decision[j, 3] < i+46){
      group[j] = score_decision[j, ]
    }else{
      group[j] = data.frame(부적격여부 = "N", 낙찰여부 = "N", score = 0)
    }
  }
  group = group %>% do.call(rbind, .)
  my_list_2[[i]] = group %>% filter(score > 0)
}
my_list_2

# my_length : 낙찰된 데이터들 중 가격점수 구간에 따른 데이터의 개수를 저장
my_length = c()
for(i in 1:25){
  my_length[i] = nrow(my_list[[i]])
}

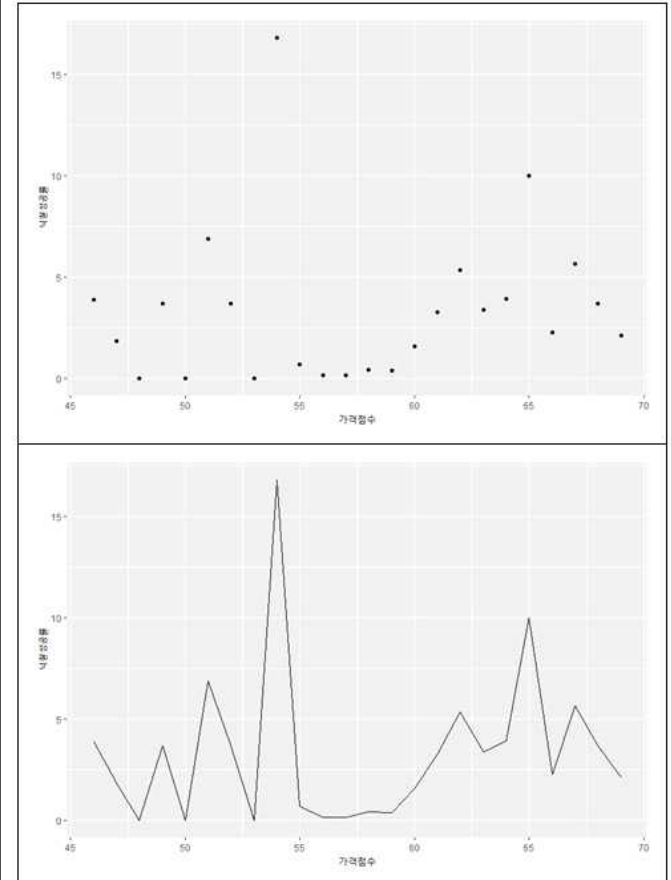
# my_length_2 : 낙찰된 여부와 관계없이 가격점수 구간에 따른 데이터의 개수를 저장
my_length_2 = c()
for(i in 1:25){
  my_length_2[i] = nrow(my_list_2[[i]])
}

# final_dat : 가격점수와 낙찰성공률(my_length와 my_length_2 이용)을 저장
final_dat = data.frame(가격점수 = c(46:69), 낙찰성공률 = c(46:69))
for(i in 1:nrow(final_dat)){
  final_dat[i, 2] = my_length[i]/my_length_2[i]
}

final_dat[, 1] = final_dat[, 1] %>% as.numeric()
summary(final_dat[, 2])

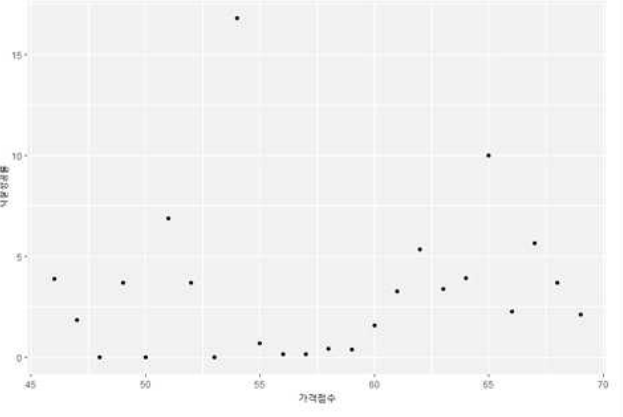
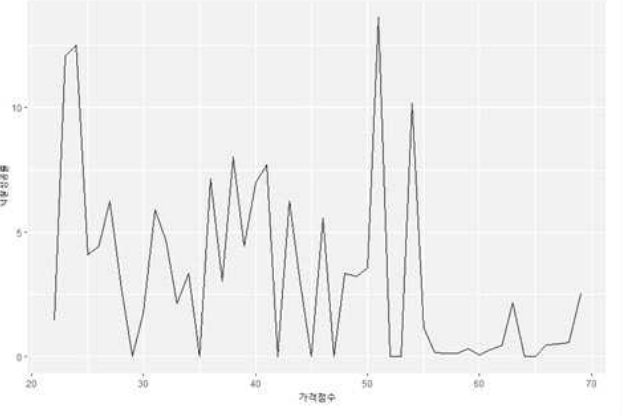
# ggplot2 시각화(가격점수와 낙찰성공률의 관계)
require(ggplot2)
ggplot(data = final_dat) + geom_line(aes(x = 가격점수, y = 낙찰성공률))
ggplot(data = final_dat) + geom_point(aes(x = 가격점수, y = 낙찰성공률))
```

3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드



3-23. Rstudio 코드.hwp

1월 공공조달 빅데이터 경진대회 활동일지 - 하반기(15일 ~ 21일)

날짜	1월 21일	방법	Meeting		
활동	가격점수와 낙찰성공률의 상관관계 코딩	멤버	정진혁, 서지영		
<div>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</div> <div># 필요한 패키지 불러오기 rm(list = ls()); gc(reset = T) require(readr); require(dplyr) # 엑셀 문서 읽기(낙찰자결정방법, 추정가격, 예정가격, 입찰금액, 입찰률, 부적격여부, 부적격사유, 낙찰여부) Excel = read_xlsx("/조달청/3-23. 물품입찰분류별 투찰업체 내역(인덱싱0).xlsx") %>% data.frame(stringsAsFactors = F) # 고시금액미만인 데이터(76683개)에서 추정가격, 입찰금액 제외한 데이터 인덱싱 criteria = grep("고시금액미만", Excel[, 1]) sub_dat = Excel[criteria, c(1, 3, 5, 6, 7, 8)] price_dat = sub_dat[, c(3, 4, 6)] price_dat = price_dat %>% mutate(score = c(1:nrow(price_dat))) for(i in 1:nrow(price_dat)){ price_dat[i, 4] = 70-4*abs(88-price_dat[i, 1]) } price_dat # score_decision : 낙찰된 것 / 안된 것 - 가격점수로 오름차순 정렬 score_decision = price_dat[, c(2, 3, 4)] score_decision = score_decision %>% arrange(score) score_decision = score_decision %>% filter(부적격여부 == "N") # 낙찰된 데이터만 가격점수 구간별로 나눔(my_list) score_decision_Yes = score_decision %>% filter(낙찰여부 == "Y") score_decision_Yes my_list = list() for(i in 1:48){ group = list() for(j in 1:nrow(score_decision_Yes)){ if(score_decision_Yes[j, 3] >= i+21 & score_decision_Yes[j, 3] < i+22){ group[[j]] = score_decision_Yes[j,] }else{ group[[j]] = data.frame(부적격여부 = "N", 낙찰여부 = "Y", score = 0) } } group = group %>% do.call(rbind, .) my_list[[i]] = group %>% filter(score > 0) } my_list</div>		<div>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</div> <div># 낙찰된 것 안된 것 모두 가격점수 구간별로 나눔(my_list_2) my_list_2 = list() for(i in 1:48){ group = list() for(j in 1:nrow(score_decision)){ if(score_decision[j, 3] >= i+21 & score_decision[j, 3] < i+22){ group[[j]] = score_decision[j,] }else{ group[[j]] = data.frame(부적격여부 = "N", 낙찰여부 = "N", score = 0) } } group = group %>% do.call(rbind, .) my_list_2[[i]] = group %>% filter(score > 0) } my_list_2 # my_length : 낙찰된 데이터들 중 가격점수 구간에 따른 데이터의 개수를 저장 my_length = c() for(i in 1:48){ my_length[i] = nrow(my_list[[i]]) } # my_length_2 : 낙찰된 여부와 관계없이 가격점수 구간에 따른 데이터의 개수를 저장 my_length_2 = c() for(i in 1:48){ my_length_2[i] = nrow(my_list_2[[i]]) } # final_dat : 가격점수와 낙찰성공률(my_length와 my_length_2 이용)을 저장 final_dat = data.frame(가격점수 = c(22:69), 낙찰성공률 = c(22:69)) for(i in 1:nrow(final_dat)){ final_dat[i, 2] = my_length[i]/my_length_2[i]*100 } final_dat[, 1] = final_dat[, 1] %>% as.numeric() summary(final_dat[, 2]) # ggplot2 시각화(가격점수와 낙찰성공률의 관계) require(ggplot2) ggplot(data = final_dat) + geom_line(aes(x = 가격점수, y = 낙찰성공률)) ggplot(data = final_dat) + geom_point(aes(x = 가격점수, y = 낙찰성공률))</div>		<div>3-23. 물품 입찰분류별 투찰업체 내역 - Rstudio 코드</div> <div> </div>	
3-23. Rstudio 코드.hwp					