

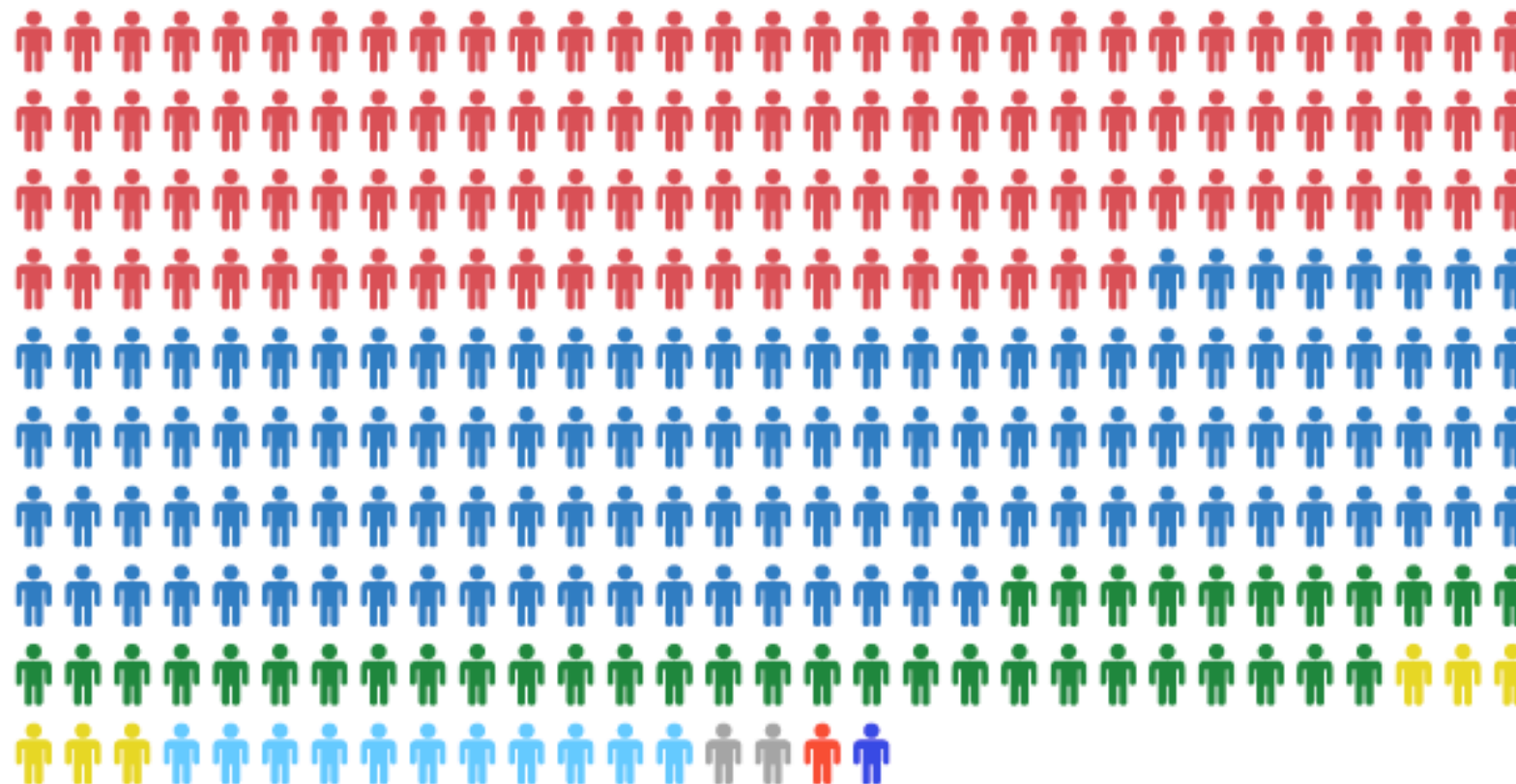
웹스크롤링을 이용한 정당별 현직 국회의원 재산 비교

2018. 1. 29

김승혁

20대 국회의원 총 297명 현황


20대 국회 정당별 의석 (2018.1.8. 기준, 단위: 명)



● 자유한국당 ● 더불어민주당 ● 국민의당 ● 정의당 ● 바른정당 ● 무소속 ● 민중당
● 대한애국당

동 사이트에서 각 의원들의 프로필 및 신고재산 현황을 볼 수 있음.
현직 국회의원 292명 중 무소속 2명, 민중당 1명, 대한애국당 1명을 제외한 288명의 신고재산 액수를 당별로 파악해보기로 함.

권은희 權垠希



● 국민의당

광주 광산구을

2선 [19대, 20대]

행정안전위원회
국회운영위원회
사법개혁특별위원회

전남대학교 법과대학

권은희 법률사무소
서울수서경찰서 수사과장

의원회관 904호 | 전화 02-784-1813 | 팩스 02-788-0307

kwoneunhee0215@gmail.com

트위터

페이스북

자료출처 : 국

대표발의 법안 총 21개

모두 입법권을 갖고 있는 헌법기관입니다. 대표발의 의원은 법률 제개정안의 기초를 마련하거나 입법을 주

5. 재산		
연도별 신고 재산 단위 : 천원		
국회의원 재산현황은 공직자윤리법에 따라 매년 한 차례 공개됩니다		
세부항목 (단위:천원)		
건물		
고지거부 및 등록제외사항		
골동품 및 예술품		
금 및 백금		
보석류		
부동산에 관한 규정이 준용되는 권리와 자동차 건설기계 선박 및 항공기		
	-1,453,301	-1,424,043
	0	0
	0	0
	0	0
	0	0
총계	1,260,328	831,062

1. 각 정당 의원들의 링크 스크롤링

- 먼저, 각 당의 페이지에서 각 의원들의 링크 주소를 스크롤링하기 위해 아래와 같은 코드를 짤.
- 각 당의 페이지 주소를 우측 아래와 같은 data frame으로 만든 뒤, for loop를 이용하여 각 당 의원들의 페이지를 스크롤링하게끔 함(전체 소스 코드는 맨 뒤 부록 참고)

```
t<-NULL
party<-c("국민의당", "더불어민주당",
        "바른정당", "자유한국당", "정의당")
addr<-c("http://watch.peoplepower21.org/?
mid=AssemblyMembers&mode=search&party=%E
A%B5%AD%EB%AF%BC%EC%9D%98%EB%8
B%B9&region=&sangim=&gender=&elect_num=&p
age=", ...생략...)
pt<-data.frame(party=party, html=addr)

h<- read_html(paste0(pt$html[a], b))
url <- html_nodes(h, css="a, a:hover") %>%
  html_attr('href')
url<-url[str_detect(url, "Member&member_seq")]
url<-paste0("http://watch.peoplepower21.org", url)
link<-NULL
for(i in 1:length(url)){
  if(i%%2==0){
    link[i/2]<-url[i] } }
```



watch.peoplepower21.org/?act=&mid=AssemblyMembers&vid=&mode=search&name=&party=국민의...

확인 검색초기화

총 297명의 의원 중 / 검색결과 39명

권은희 김경진 김관영 김광수
김동철 김삼화 김성식 김수민
김종희 김종로 박선숙 박주선

	party	html
1	국민의당	http://watch.peoplepower21.org/?mid=AssemblyMembers...
2	더불어민주당	http://watch.peoplepower21.org/?mid=AssemblyMembers...
3	바른정당	http://watch.peoplepower21.org/?act=&mid=AssemblyMe...
4	자유한국당	http://watch.peoplepower21.org/?mid=AssemblyMembers...
5	정의당	http://watch.peoplepower21.org/?act=&mid=AssemblyMe...

(전체 소스 코드는 맨 뒤 부록 참고)

2. 각 정당 의원들의 이름 스크롤링

- 각 당의 페이지에서 각 의원들의 이름을 스크롤링하기 위해 아래와 같은 코드를 짤.

```
c<-length(link)
if(c!=0){
  txt0 <- html_nodes(h,css="a") %>%
    html_text()
  txt0<-txt0[str_detect(txt0, "●")]
  name<-unlist(str_extract_all(txt0, "[가-
  할]{2,4}"))
```



(전체 소스 코드는 맨 뒤 부록 참고)

3. 각 정당 의원들의 재산 스크롤링

- 각 의원들의 페이지에서 각 의원들의 재산을 스크롤링하기 위해 아래와 같은 코드를 짤.
- (각 의원들 신고재산 중 2017년의 총계만 스크롤링함)

```
money<- NULL
for(i in 1:length(link)){
  h2<- read_html(link[i])
  txt2<-html_nodes(h2, ".info")%>%
    html_text()
  if(length(txt2)>0){
    if(str_detect(txt2, ".*[0-9]{3}
[0-9]+.*[0-9]+.*")){
      money[i]<-sub(".*[0-9]{3}
([0-9]+.*[0-9]+).*$", "\\1", txt2)
    } else {
      money[i]<-NA } } }
```

5. 재산			자료출처 : 국회공보
연도별 신고 재산 단위 : 천원			
국회의원 재산현황은 공직자윤리법에 따라 매년 한 차례 공개됩니다.			
세부항목 (단위:천원)	2016	2017	
	0	0	
건물	2,453,885	1,982,211	
고지거부 및 등록제외사항	0	0	
골동품 및 예술품	0	0	
금 및 백금	0	0	
보석류	0	0	
부동산에 관한 규정이 준용되는 권리와 자동차 건설기계 선박 및 항공기	31,470	23,900	
비영리법인에의한 출연재산	0	0	
예금	75,512	70,055	
유가증권	140,000	140,000	
정치자금법에 따른 정치자금의 수입 및 지출을 위한 예금계좌의 예금	12,762	38,939	
지식재산권	0	0	
채권	0	0	
채무	-1,453,301	-1,424,043	
토지	0	0	
합영 합자 유한회사 출자지분	0	0	
현금	0	0	
회원권	0	0	
총계	1,260,328	831,062	

(전체 소스 코드는 맨 뒤 부록 참고)

4. 정당이름, 의원이름, 재산을 컬럼으로 data frame 생성

- 앞에서 만든 코드와 while 및 for 루프를 이
용해서 모든 의원들의 데이터를 rbind로 t
라는 데이터프레임에 넣음.
- 데이터프레임이 만들어진 결과를 확인하면
오른쪽과 같음.
- 재산을 숫자 형식으로 만들어 계산하기 위
해 gsub 및 as.numeric함수로 변환.

```
for(a in 1:5){  
  b<-1  
  c<-1  
  while(c!=0){  
    h<- read_html(paste0(pt$html[a], b))  
    ...중략...  
    b<-b+1  
    c<-length(link)  
    ...중략...  
    t<-rbind(t, data.frame(party=pt$party[a],  
name=name, money=money,  
stringsAsFactors = F)) } }  
  
t$money<-gsub(",", "", t$money)  
t$money <- as.numeric(t$money)
```

View(t)

	party	name	money
25	국민의당	이용주	1,627,720
26	국민의당	이용호	1,510,225
27	국민의당	이찬열	939,846
28	국민의당	이태규	644,517
29	국민의당	장병완	7,604,065
30	국민의당	장정숙	141,241
31	국민의당	정동영	1,140,172
32	국민의당	정인화	1,582,571
33	국민의당	조배숙	1,767,554
34	국민의당	주승용	5,088,998
35	국민의당	채이배	537,739
36	국민의당	천정배	772,522
37	국민의당	최경환	230,028
38	국민의당	최도자	633,844
39	국민의당	황주홍	2,785,736
40	더불어민주당	강병원	754,761
41	더불어민주당	강창일	2,111,355

(전체 소스 코드는 맨 뒤 부록 참고)

스크롤링 결과 검증

View(aggregate(name~party, t, length))

- 각 정당 의원들 데이터가 모두 들어간 것을 확인함.

	party	name
1	국민의당	39
2	더불어민주당	121
3	바른정당	9
4	자유한국당	118
5	정의당	6

결측값 처리

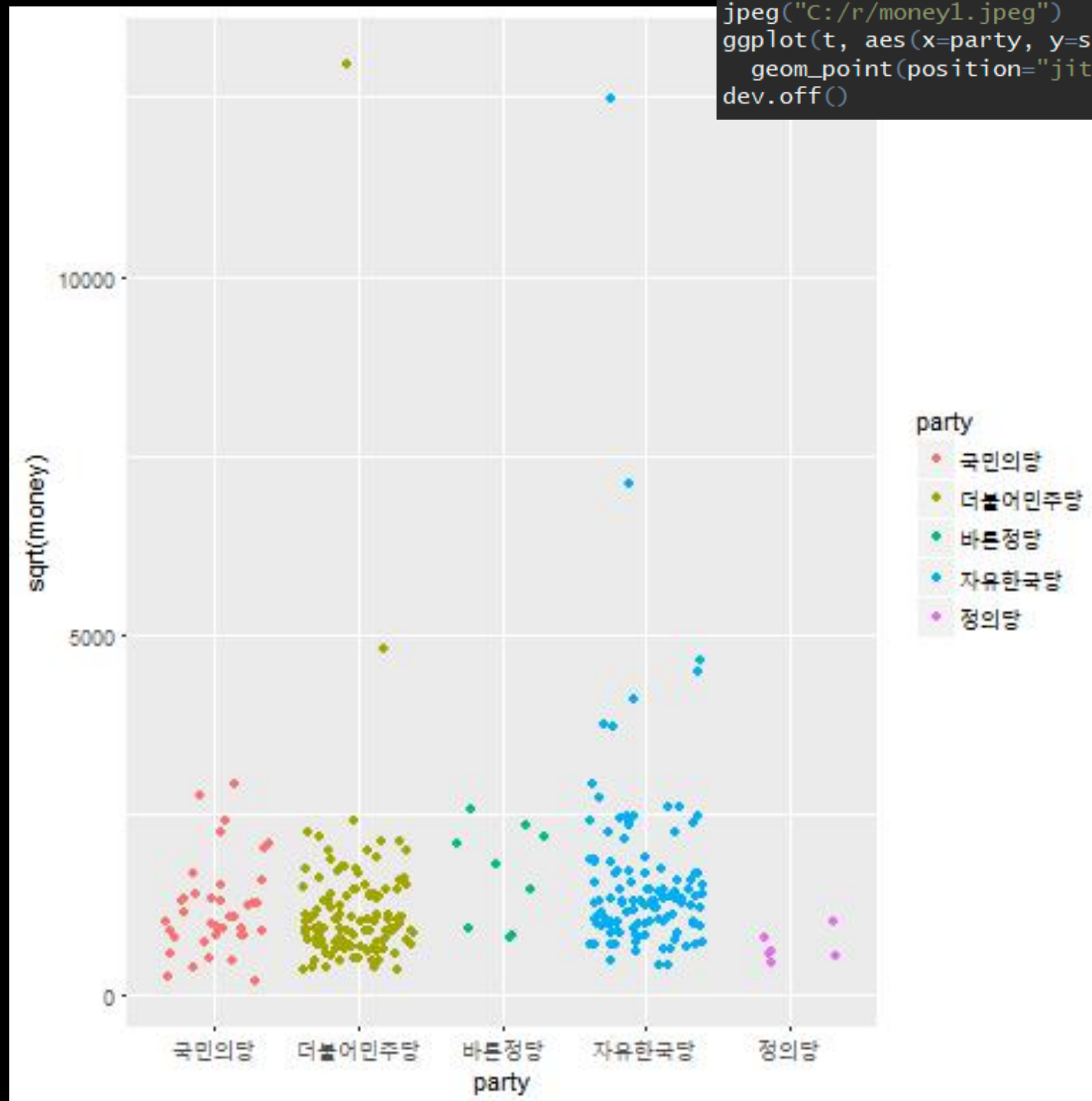
- 수집된 결과(data frame)의 재산 항목이 NA인 경우는 총 6건으로, 더불어민주당 3건, 자유한국당 3건이었음.
- 해당 웹사이트에 기재된 재산이 0원인 경우(심기준 의원), 마이너스인 경우(진선미 의원, 김한표 의원), 신고된 재산이 없는 경우(이수혁 의원, 김성태 의원, 김재원 의원)
- 해당 결측값들은 제외시켜도 더불어민주당, 자유한국당 총 의원의 재산 분포를 파악하는 데에 문제가 되지 않는다고 봐서 제외시킴.

View(t[is.na(t\$money),])

	party	name	money
104	더불어민주당	심기준	NA
126	더불어민주당	이수혁	NA
150	더불어민주당	진선미	NA
191	자유한국당	김성태	NA
198	자유한국당	김재원	NA
205	자유한국당	김한표	NA

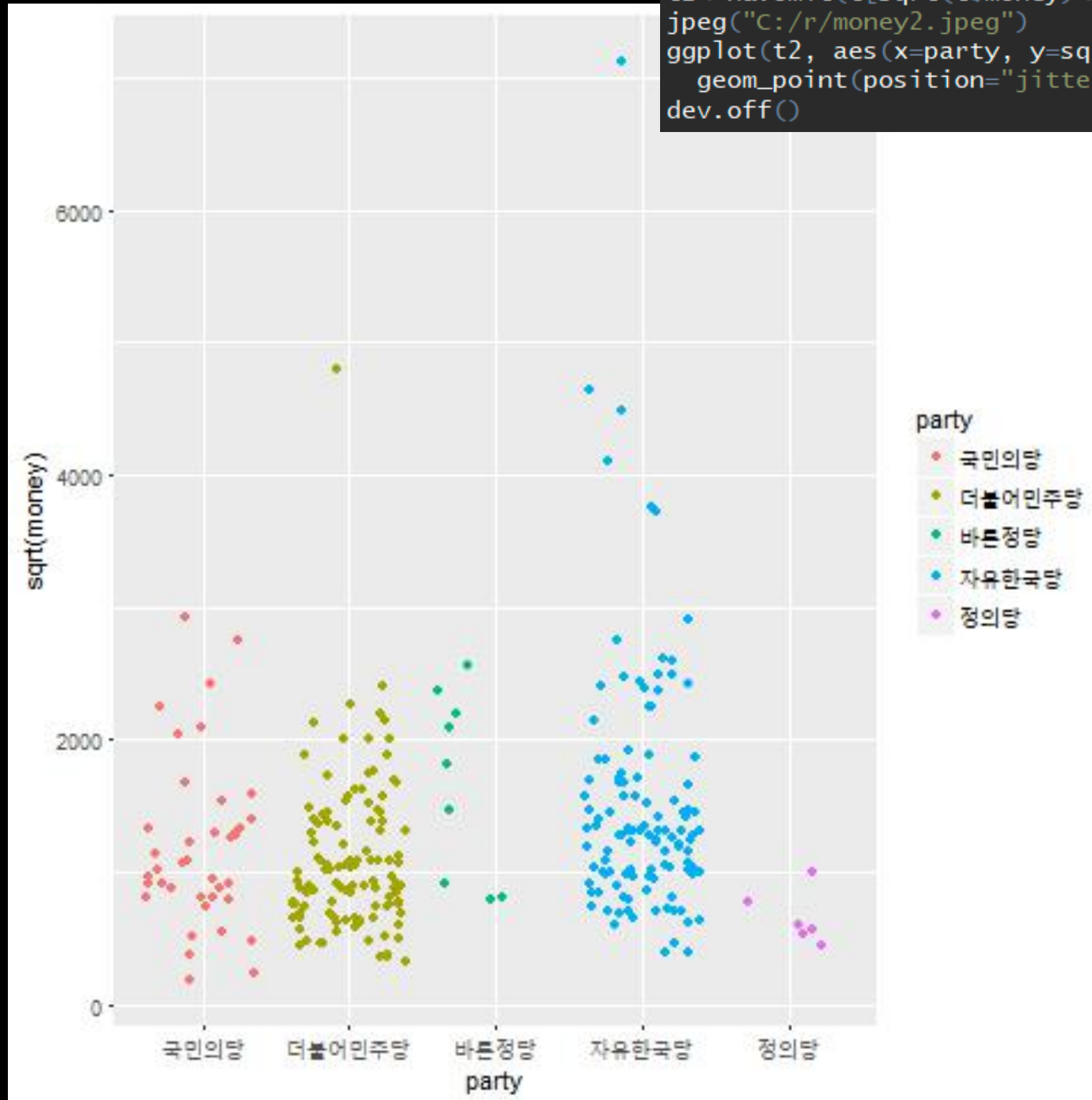
첫 결과물.

특이치(재산이 1천억원이 넘는 경우 2건) 2건으로 인해 전체적인 분포 파악에 방해가 됨.
(더불어민주당 김병관의원, 자유한국당 김세연의원)

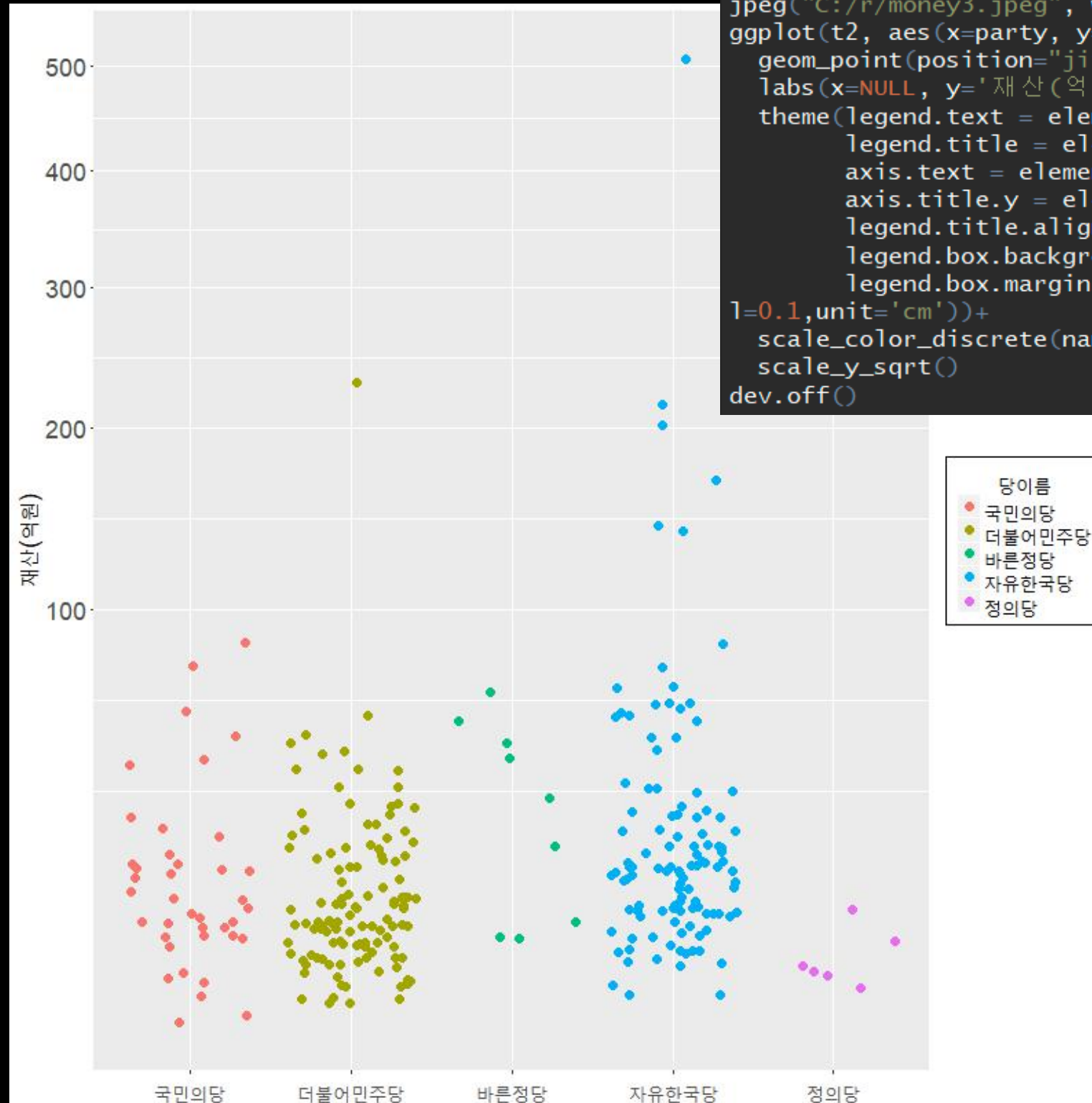


재산이 1천억원이 넘는 값(더불어민주당 김병관의원, 자유한국당 김세연의원)을 버린 2번째 결과물.
각 당의 전체적인 재산 분포를 볼 수 있음. 다만 Y축이 한 눈에 보기 어려우므로 다음 결과물에서 Y축 수정.

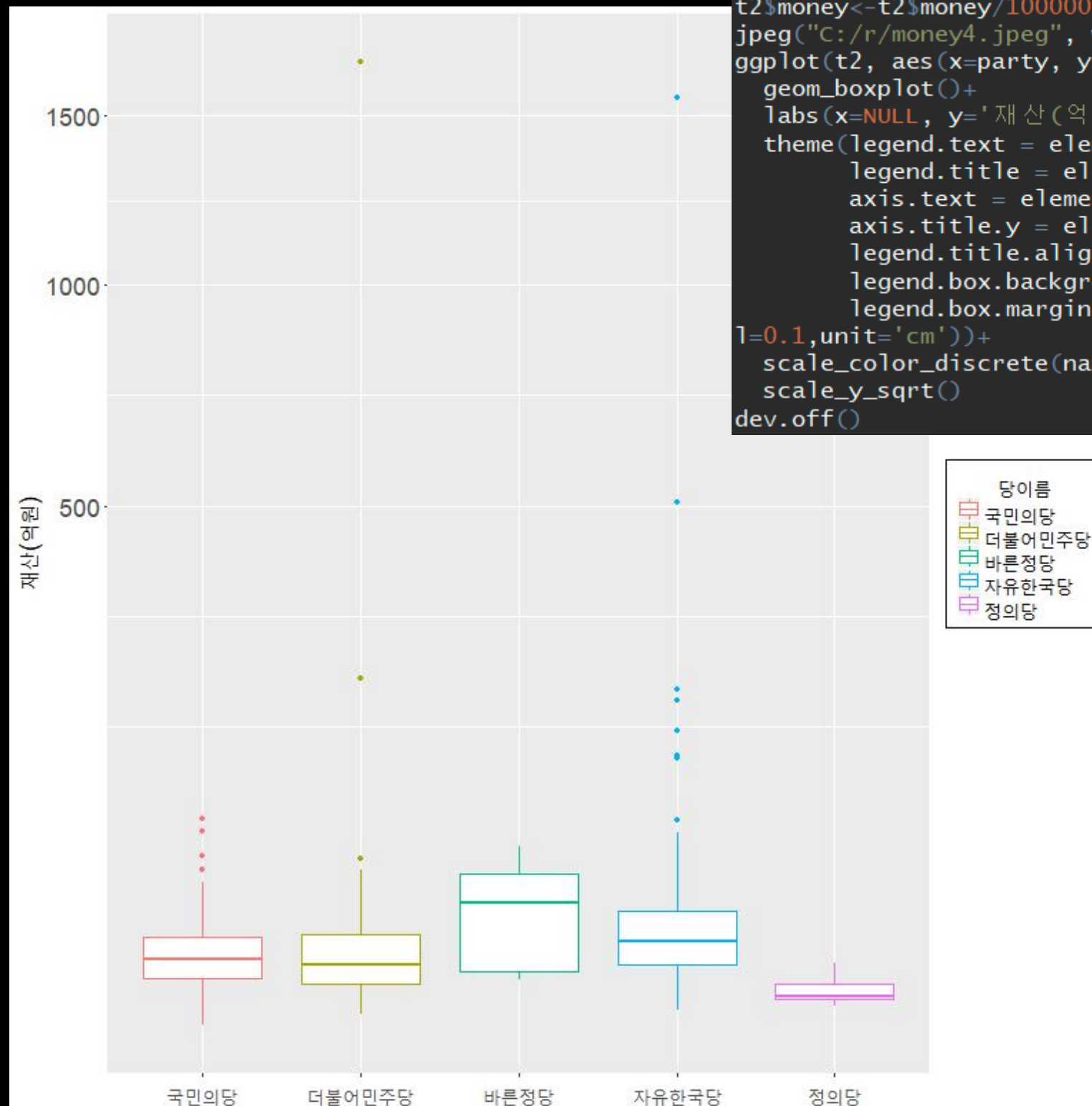
```
t2<-na.omit(t[sqrt(t$money)<10000,])  
jpeg("C:/r/money2.jpeg")  
ggplot(t2, aes(x=party, y=sqrt(money), colour=party))+  
  geom_point(position="jitter")  
dev.off()
```



최종 결과물. Y축 이름을 재산(100만원 단위)으로 수정하고,
값의 분포는 ggplot의 `scale_y_sqrt()` 메소드를 이용하여 보기 쉽게 나타냄.



최종 결과물 2번째. Box plot 을 이용하여 종위수, 사분위수, 특이치까지 나타냄.



```
t2<-na.omit(t)
t2$money<-t2$money/100000
jpeg("C:/r/money4.jpeg", width = 800, height = 800)
ggplot(t2, aes(x=party, y=money, colour=party))+
  geom_boxplot()+
  labs(x=NULL, y='재 산(억 원)')+
  theme(legend.text = element_text(size=18),
        legend.title = element_text(size=18),
        axis.text = element_text(size=18),
        axis.title.y = element_text(size=18),
        legend.title.align=0.5,
        legend.box.background = element_rect(),
        legend.box.margin = margin(t=0.1,r=0.1, b=0.1,
l=0.1,unit='cm'))+
  scale_color_discrete(name="당 이 름")+
  scale_y_sqrt()
dev.off()
```

사용한 패키지

- `library(stringr)`
- `library(rvest)`
- `library(dplyr)`
- `library(ggplot2)`
- `library(jpeg)`

전체 소스 코드 (1)

```
## {r}
library(stringr)
library(rvest)
library(dplyr)
library(ggplot2)
library(jpeg)
t<-NULL
party<-c("국민의당", "더불어민주당",
        "바른정당", "자유한국당", "정의당")
addr<-c("http://watch.peoplepower21.org/?mid=AssemblyMembers&mode=search&party=%EA%B5%AD%EB%AF%BC%EC%9D%98%EB%8B%B9&region=&sangim=&gender=&elect_num=&page=",
        "http://watch.peoplepower21.org/?mid=AssemblyMembers&mode=search&party=%EB%8D%94%EB%B6%88%EC%96%B4%EB%AF%BC%EC%A3%BC%EB%8B%B9&region=&sangim=&gender=&elect_num=&page=",
        "http://watch.peoplepower21.org/?act=&mid=AssemblyMembers&vid=&mode=search&name=&party=%EB%B0%94%EB%A5%B8%EC%A0%95%EB%8B%B9&region=&sangim=&gender=&age=&elect_num=&page=",
        "http://watch.peoplepower21.org/?mid=AssemblyMembers&mode=search&party=%EC%9E%90%EC%9C%A0%ED%95%9C%EA%B5%AD%EB%8B%B9&region=&sangim=&gender=&elect_num=&page=",
        "http://watch.peoplepower21.org/?act=&mid=AssemblyMembers&vid=&mode=search&name=&party=%EC%A0%95%EC%9D%98%EB%8B%B9&region=&sangim=&gender=&age=&elect_num=&page=")
pt<-data.frame(party=party, html=addr)
```


전체 소스 코드 (2)

```
for(a in 1:5){
  b<-1
  c<-1
  while(c!=0){
    h<- read_html(paste0(pt$html[a], b))
    url <- html_nodes(h, css="a, a:hover") %>%
      html_attr('href')
    url<-url[str_detect(url, "Member&member_seq")]
    url<-paste0("http://watch.peopower21.org", url)
    link<-NULL
    for(i in 1:length(url)){
      if(i%%2==0){
        link[i/2]<-url[i]
      }
    }
    b<-b+1
    c<-length(link)
    if(c!=0){
      txt0 <- html_nodes(h,css="a") %>%
        html_text()
      txt0<-txt0[str_detect(txt0, "●")]
      name<-unlist(str_extract_all(txt0, "[가-힣]{2,4}"))
      money<- NULL
    }
  }
}
```


전체 소스 코드 (3)

```
for(i in 1:length(link)){
h2<- read_html(link[i])
txt2<-html_nodes(h2, ".info")%>%
  html_text()
if(length(txt2)>0){
if(str_detect(txt2, ".*,[0-9]{3}[0-9]+.*[0-9]+.*")){
  money[i]<-sub(".*,[0-9]{3}([0-9]+.*[0-9]+).*$", "\\1", txt2)
}else{
  money[i]<-NA
}
}
}
t<-rbind(t, data.frame(party=pt$party[a], name=name,
money=money, stringsAsFactors = F))
}
}
}

t$money<-gsub(",", "", t$money)
t$money <- as.numeric(t$money)
```

전체 소스 코드 (4)

```
jpeg("C:/r/money1.jpeg")
ggplot(t, aes(x=party, y=sqrt(money), colour=party))+
  geom_point(position="jitter")
dev.off()

t2<-na.omit(t[sqrt(t$money)<10000,])
jpeg("C:/r/money2.jpeg")
ggplot(t2, aes(x=party, y=sqrt(money), colour=party))+
  geom_point(position="jitter")
dev.off()
```

전체 소스 코드 (5)

```
t2$money<-t2$money/100000
jpeg("C:/r/money3.jpeg", width = 800, height = 800)
ggplot(t2, aes(x=party, y=money, colour=party))+
  geom_point(position="jitter", size = 3)+
  labs(x=NULL, y='재산(억 원)')+
  theme(legend.text = element_text(size=18),
        legend.title = element_text(size=18),
        axis.text = element_text(size=18),
        axis.title.y = element_text(size=18),
        legend.title.align=0.5,
        legend.box.background = element_rect(),
        legend.box.margin = margin(t=0.1,r=0.1, b=0.1,
l=0.1,unit='cm'))+
  scale_color_discrete(name="당 이 름")+
  scale_y_sqrt()
dev.off()
```

전체 소스 코드 (6)

```
t2<-na.omit(t)
t2$money<-t2$money/100000
jpeg("C:/r/money4.jpeg", width = 800, height = 800)
ggplot(t2, aes(x=party, y=money, colour=party))+
  geom_boxplot()+
  labs(x=NULL, y='재 산(억 원)')+
  theme(legend.text = element_text(size=18),
        legend.title = element_text(size=18),
        axis.text = element_text(size=18),
        axis.title.y = element_text(size=18),
        legend.title.align=0.5,
        legend.box.background = element_rect(),
        legend.box.margin = margin(t=0.1,r=0.1, b=0.1,
l=0.1,unit='cm'))+
  scale_color_discrete(name="당 이 름")+
  scale_y_sqrt()
dev.off()
```