따릉이 데이터 분석

중앙대학교 물리학과
20142927 박원빈

1. 소개

서울시에서 2015년 9월부터 공공자전거 사업(따릉이)을 시작했다. 지난 5년간 따릉이를 사용하면서 수리가 필요한 자전거를 많이 봤다. 따릉이 사업 소개에 따르면 잔고장이 안 나는 내구성이 강한 설계로 제작된 모델이라고 한다. 과연 얼마나 튼튼한지 확인해보고 싶어 연구를 시작했다. 이 연구는 서울시 공공데이터에 올라온 따릉이 고장 신고 내역과 자전거 대여 내역과 데이터를 사용한다. 이 연구는 1. 계절에 따른 자전거 사용과 고장 신고 수의수 의 관계와 2. 고장접수가 많이 된 자전거(고유번호)의 누적사용량(사용 시간, 이동거리)을 시각화하고 회귀 식을 도출한다.

데이터 전처리

서울시 공공데이터 자료실에 게시된 raw-data에 사소한 문제가 있었다. 2020.1월 ~ 2020.5월 파일에 한글 인코딩은 중간중간 깨져 데이터 열 구분이 틀어져있는 경우가 종종 있다. 그리고 대여소 ID를 기록하는 열은 정수형 데이터가 예상되지만 가끔씩 대여 소명 (와트컴, 중량센터, 상암센터 등)이 들어가 있기도 하다. 해당 이슈를 보정하는 작업은 첨부파일과 깃허브에 정리했다.

- 깃허브 주소 : https://github.com/ParkWonBin/R_Seoul_Bike_DataAnalisys
- 대여소 목록 : https://data.seoul.go.kr/dataList/OA-13252/F/1/datasetView.do
- 따릉이 고장 : https://data.seoul.go.kr/dataList/OA-15644/F/1/datasetView.do
- 따릉이 사용내역 : https://data.seoul.go.kr/dataList/OA-15182/F/1/datasetView.do

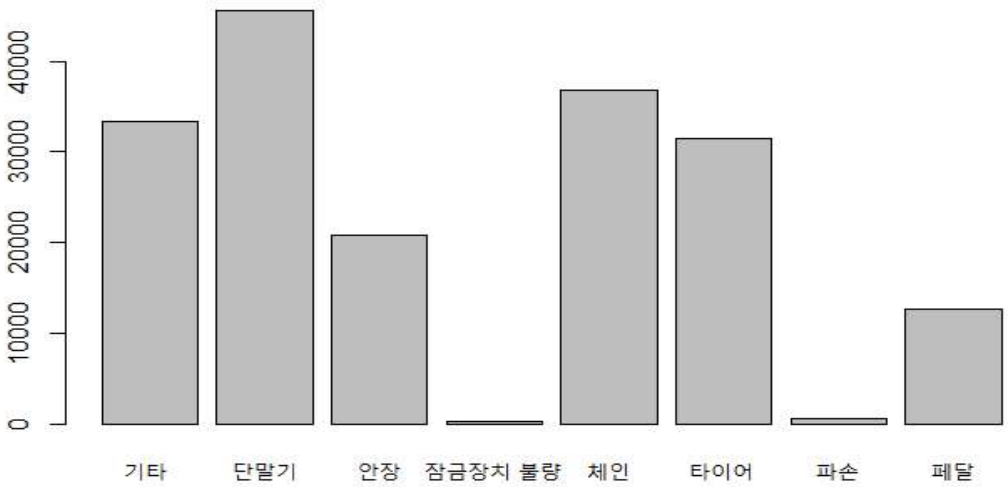
공공데이터를 모두 다운받으면 5.3GB 정도가 된다. 하지만 데이터 내에 중복되는 데이터와 문자열이 많다. 문자열로 자전거 ID를 정수형으로 바꾸고, 문자열로 기록되 날짜 데이터를 정수형 데이터로 변환(게시일 2015.9.18로부터 +n일)하여 다시 저장했다. 필요한 정보만 남기니 5.3GB(파일 52개)가 1.2GB csv파일 하나로 정리됐다. csv를 zip파일로 압축하니 358.5MB로 압축이 되었다. 이 문서는 전처리가 완료된 이후 코드부터 다룬다.

2. 데이터 둘러보기

2.1. 잘 고장나는 부품

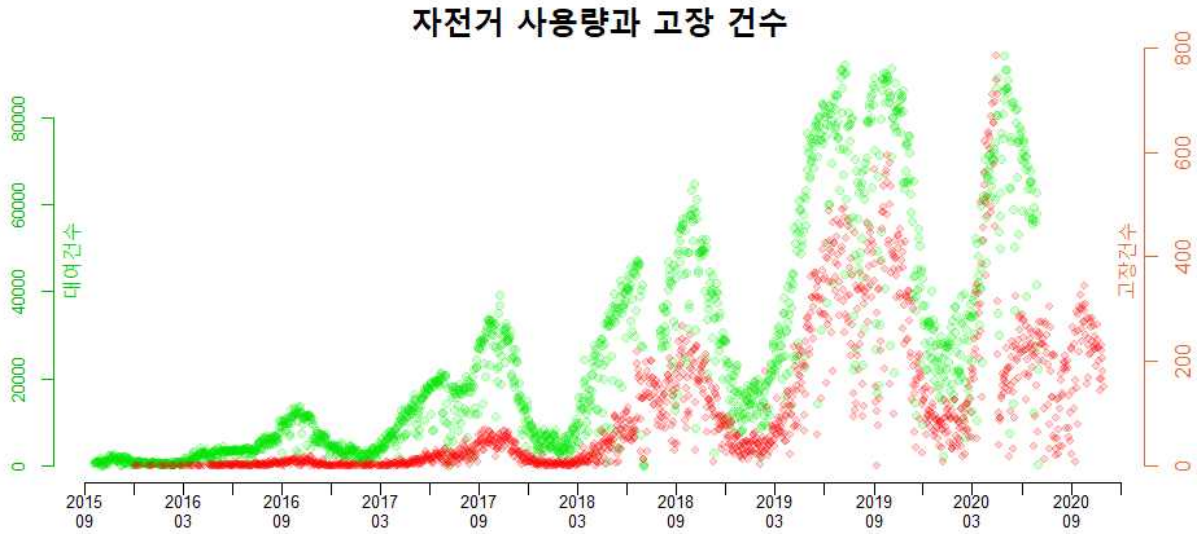
가장 고장이 많은 부품은 **1.단말기, 2.체인 3.기타 4.타이어** 순이다. 2020년 3월 1일 이후에 도입한 뉴 따릉이(QR코드형)은 자전거에서 단말기를 제거되었다. 앞으로는 단말기로 인한 고장은 사라질 예정이다.

기타.	단말기	안장	잠금장치.불량	체인	타이어.	파손	펑달
33356	45613	20785	251	36870	31458	629	12630



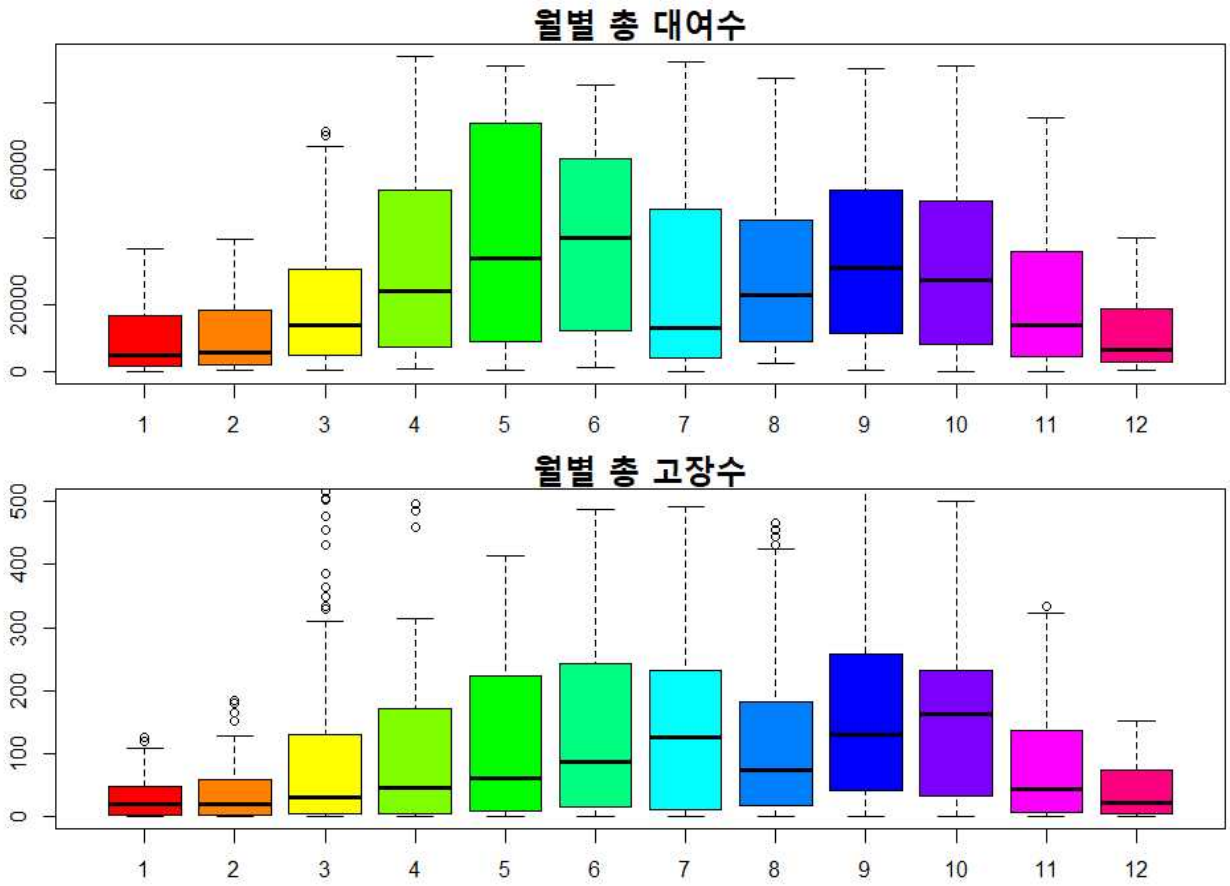
2.2. 고장나기 좋은 계절

자전거 사용은 계절성을 나타낸다. 자전거 사용의 계절성과 자전거 고장의 계절성에 대해서 시각화 하려고 한다. 직관적으로 사람들이 자전거를 많이 사용하는 때 고장 신고도 많이 발생한다는 것을 생각할 수 있다. 이 직관이 통계적으로 얼마나 믿을만한지 분석해보려고 한다.



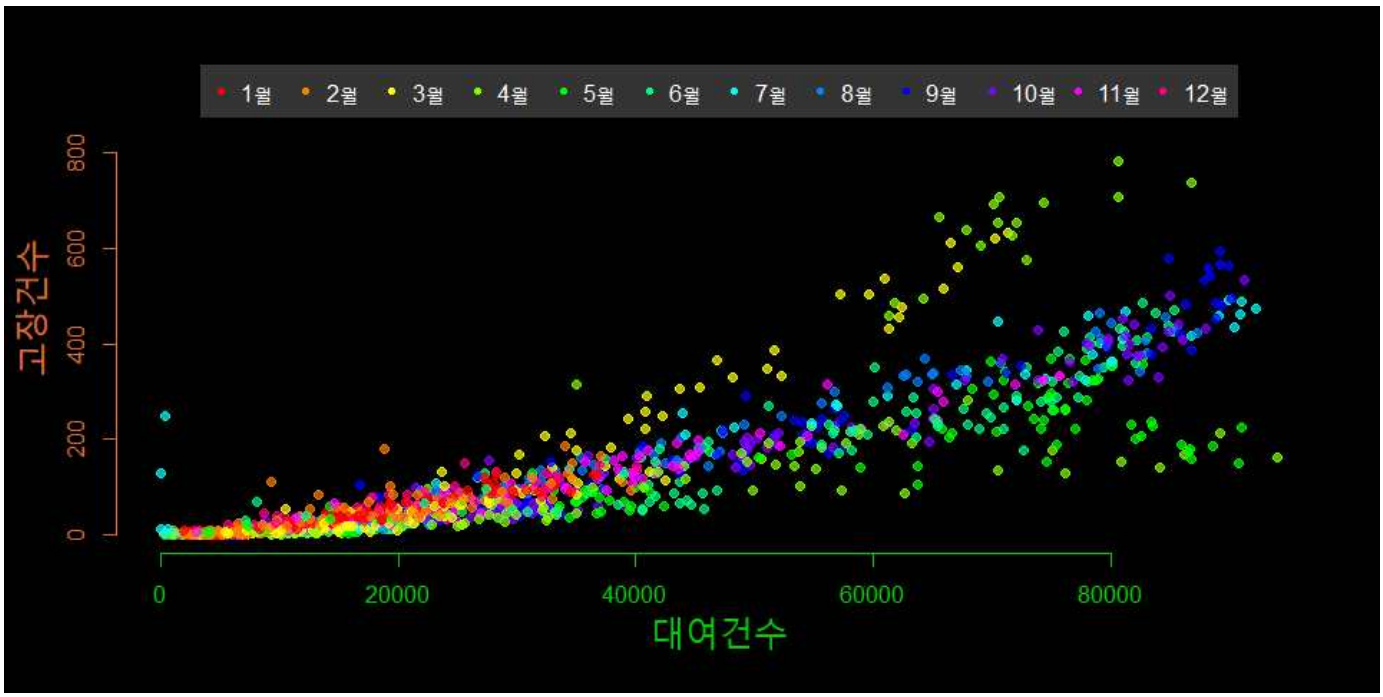
따릉이 자전거 사용량(일일 대여건수)과 고장 신고 건수는 명확히 일치하는 계절성을 보인다.
그리고 고장 건수와 대여 건수의 축을 교차해보니 '고장 건수'와 '대여 건수' 사이에 선형적인 관계가 있을 것 같은 느낌이 든다.

1. 고장 건수와 대여 건수 사이의 선형회귀하면 설명력은 어느 정도가 될까?
2. 위에서 구한 회귀식이 계절성의 영향을 받을까?

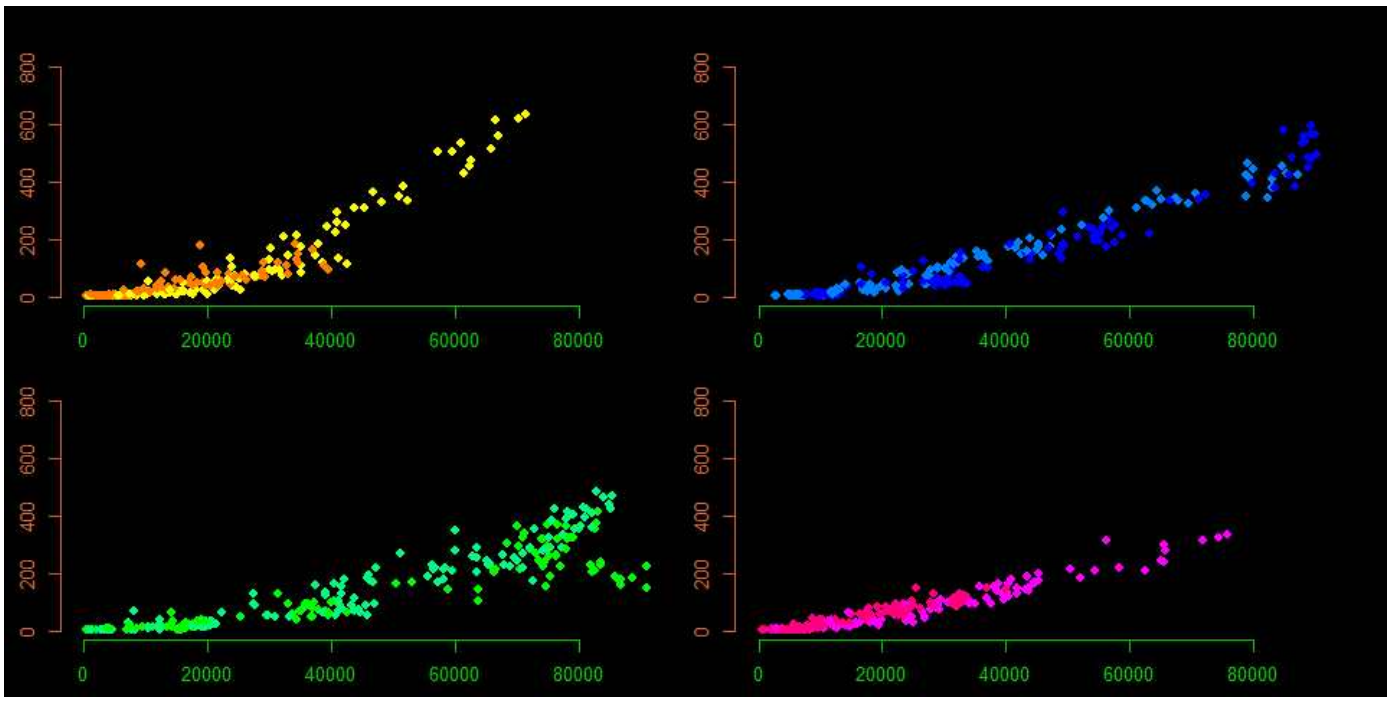


2.3. 선형성 시각화

변수 간의 선형-상관성을 시각적으로 확인하기 위해 x, y축을 고장 건수와 대여 건수로 나타냈다.
데이터가 계절성을 갖고 있기 때문에, 계절성에 대한 직관을 잃지 않고자 월별로 data의 색깔을 다르게 칠했다.
아래 그림을 살펴보면, 고장 건수와 대여 건수 사이의 기울기가 계절마다 다른 느낌이 든다.



3월(노란색) 데이터는 기울기가 가파르고 6월(초록색) 데이터는 기울기가 원만하다.
그림 하나에 데이터 전부 넣으니 분류(월)별 데이터의 경향성을 파악하기 쉽지 않다.
깔끔하게 분기별로 3달씩 그래프를 다시 그렸다.



2015~2020, 5년간의 데이터가 모인 자료인데 년 도가 달라도 월별로 비슷한 경향성을 보이는 것이 신기하다. 그림이 너무 잘 뽑혀서 2015.9.18부터 날짜별로 점이 하나씩 추가하는 그래프 애니메이션을 만들면 볼만하겠다는 생각이 든다. 하지만 과제 마감이 멀지 않아 사치 부릴 생각은 접어둔다. 회귀식이나 통계와 관련된 숫자만 구하고 다음으로 넘어갈 예정이다.

2.4 계절성에 의한 회귀식 구하기

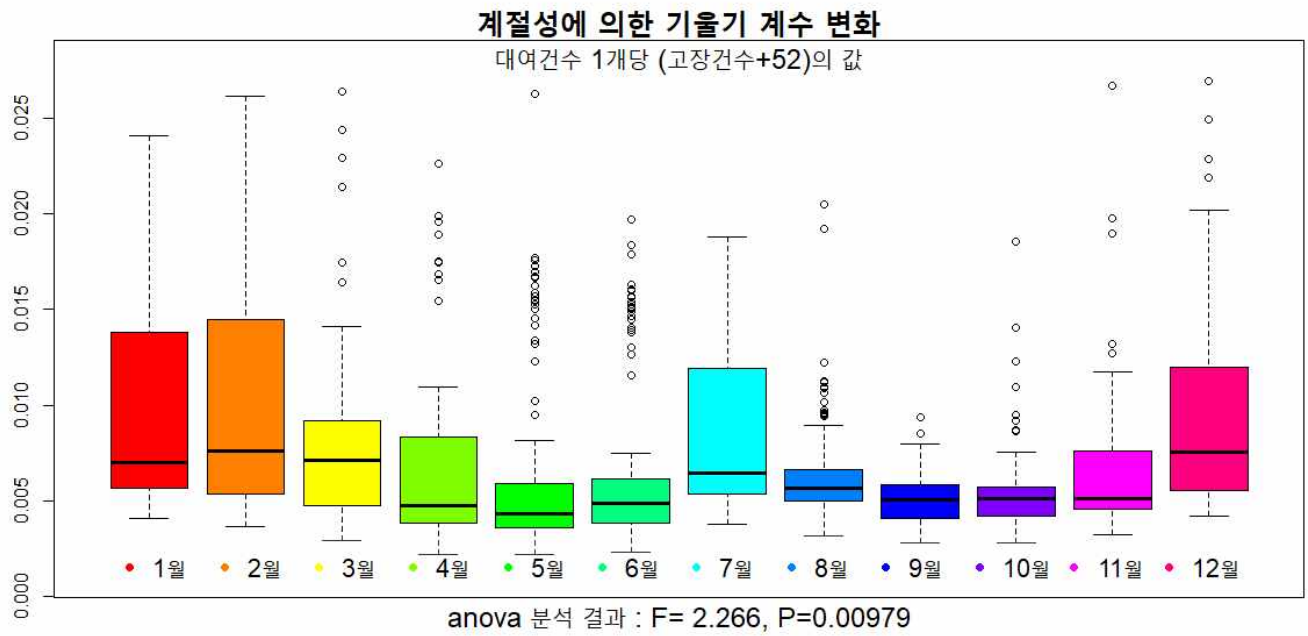
분기별로 회귀식을 돌려봤다. 데이터가 많아서 그런지 p-value가 매우 작다. 역시 GB 단위의 CSV는 믿고 쓸 수 있는 것 같다. 아래 회귀식을 통해 '자전거 대여가 많은 달에 자전거 고장 신고도 많이 들어온다.'라는 당연한 사실을 확인할 수 있다. 생각보다 일이 잘 풀리는 기분이라 작업하는 보람이 있다.

고장건수 = lm.coff*대여건수 + Intercept + error

season <chr>	lm.coff <dbl>	Intercept <dbl>	p.lmcoff <dbl>	p.intercept <dbl>	Std.error <dbl>
1~3월	0.006569172	-47.81317	5.897309e-123	1.235849e-28	3.930649
4~6월	0.004603475	-43.23544	4.474368e-105	9.241403e-09	7.375786
7~9월	0.005550117	-51.30241	3.894221e-185	1.915315e-33	3.796181
10~12월	0.004813913	-37.46420	2.316252e-206	4.030818e-40	2.486500
전체기간	0.004940014	-38.20736	0.000000e+00	6.538801e-49	2.503889

5 rows

계절별로 유도된 선형 계수(lm.coff)의 크기가 다르다. 이 계수가 계절성의 영향을 받아 변화하는 것이 맞는지 분산분석을 통해 확인해본다. 확인 방법은 위에서 얻은 회귀식을 변형하여 데이터마다 (lm.coff)값에 대응하는 수치를 계산하여 월별로 분포를 확인한다.



p-value가 작게 나와서 마음이 편하다. 회귀 식은 7월~9월 회귀 식을 기준으로, y 절편(약-52)만큼 평행이동 후 데이터마다 기울기를 계산한 것이다. 분산분석 결과 99% 유의수준에서 '기울기 데이터는 월별로 차이가 없다.'라는 귀무가설을 기각할 수 있다.

여담으로 회귀식의 y 절편을 크게 잡을수록 (원점에서 멀리 이동할수록) p 값이 낮게 나온다. 가령 원점을 y축으로 10^7만큼 이동하여 기울기를 계산하면 p 값으로 0.0001을 얻을 수 있다. 이는 데이터의 x축(대여 건수)이 계절성의 영향을 많이 받기 때문이다.

결과적으로 선형계수는 계절성의 영향을 받고, 그 주기는 6개월 정도 되어 보인다. 푸리에 변환으로 위의 그래프를 잘 설명하는 근사식 a(t) = [a*exp(iwt)+b]를 상정하면 '계절성의 영향을 보정한' 회귀 식을 구할 수 있을 것이다.

$$y = a(t)x + Intersect$$

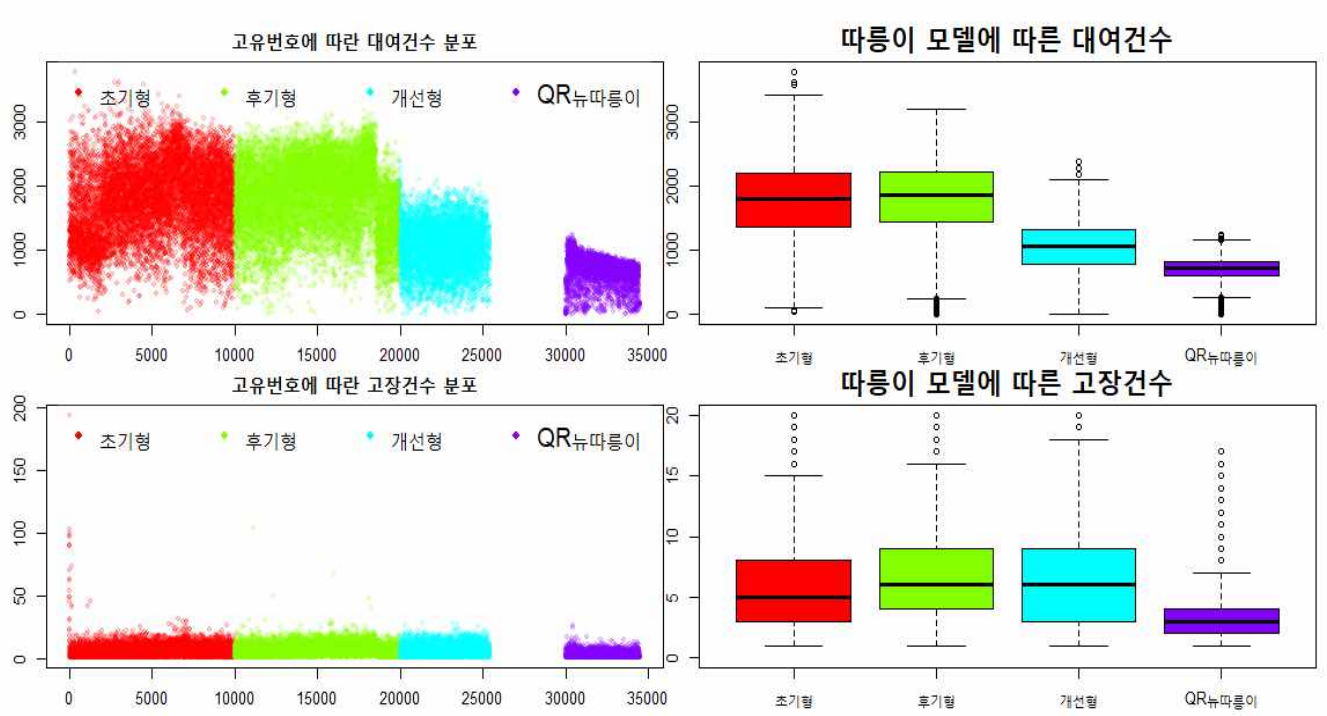
$$y = [a*exp(iwt)+b]\ x + Intersect$$

3. 고장 난 자전거(고유번호) 분석

따릉이는 자전거의 모델(종류)마다 고유번호가 다르다. 자전거 고유번호를 기준으로 자전거 모델별로 고장률에 미치는 영향을 확인하고, 자전거의 고유번호(ID)를 key 값으로 잔고장이 많은 자전거의 데이터를 분석해보려고 한다. 고장 건수, 대여 건수, 대여 시간, 이동 거리 등을 자전거 고유 번호를 기준으로 합계하여 데이터를 준비한다.

3.1. 자전거의 사용량과 고장 신고 건수

자전거의 고유번호를 x축으로 총 대여 건수와 고장 건수의 그래프를 그렸다.
모델마다 시행 일자가 달라 사용량에서는 평균의 차이는 있어보인다. box-plot과 ANOVA 분석을 한다.



따릉이 모델 간 대여건수 : F = 7024 | P <= 2e-16 ***
따릉이 모델 간 고장건수 : F = 677.4 | P <= 2e-16 ***

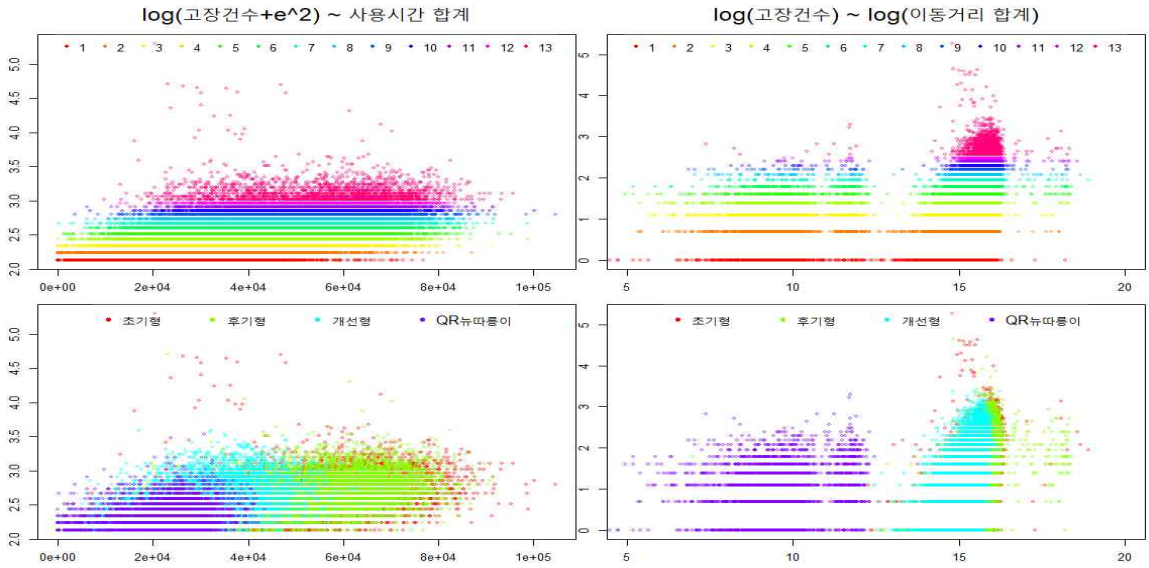
따릉이 모델에 따른 사용량과 사용량과 고장 수는 관계가 있어보인다. 분산분석 결과 매우 높은 유의수준 (1 - 2e-16)으로 집단 간 평균의 차이가 있다는 것을 알 수 있다. 데이터가 많아 전반적으로 모든 분석에 유의수준이 높게 나오는 것 같다.

3.2. 고장 횟수와 사용량(사용시간,이동거리)

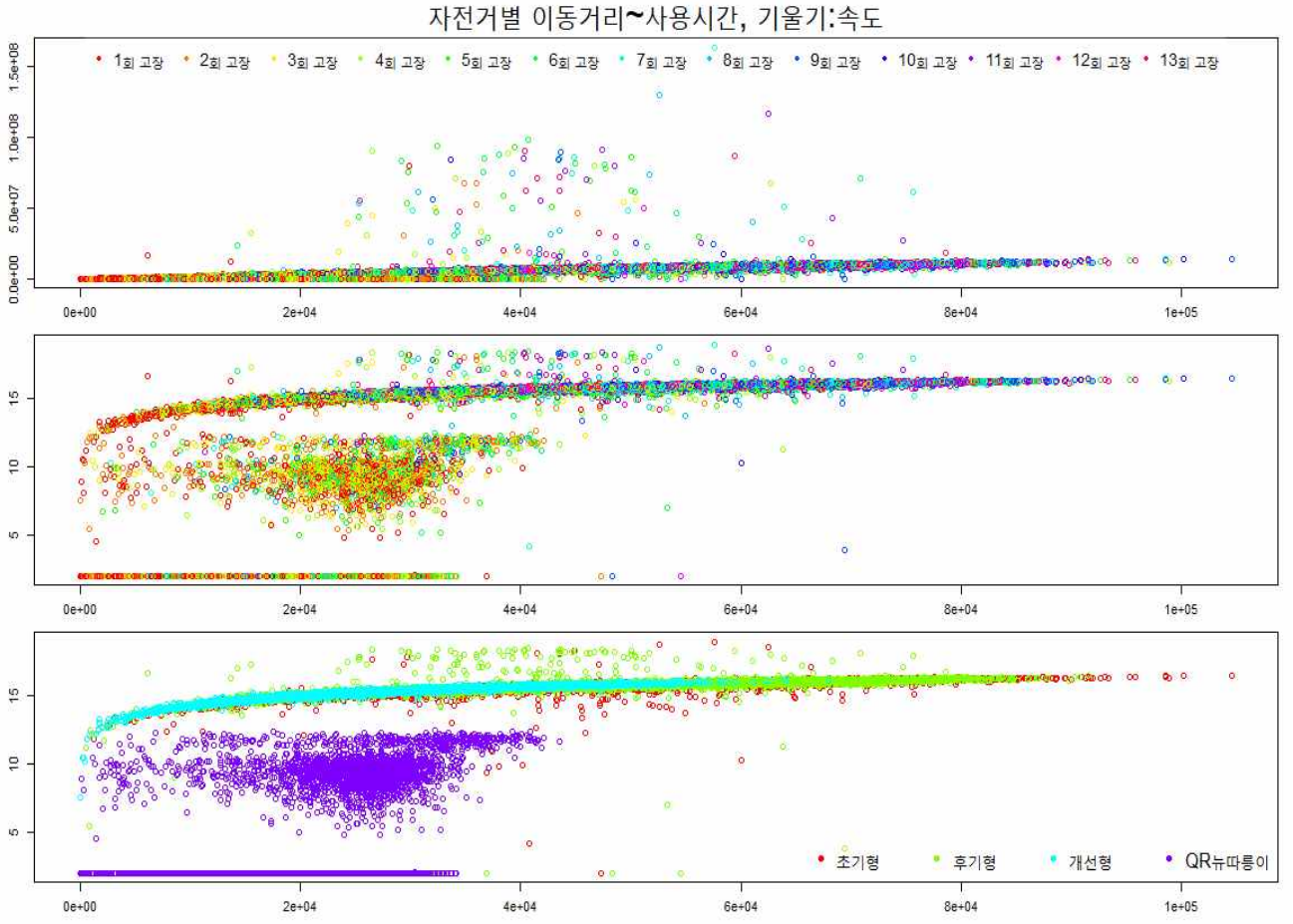
자전거 총 이용시간과 이동거리가 고장신고 수에 미치는 영향을 알아보기 위해 그래프를 그렸다. 데이터마다 사용시간과 이동거리 값의 분산이 매우 큰데, 이는 자전거마다 운행 기간이 다르고, 주로 배치된 지역의 자전거 사용률도 다르기 때문이다. 고장 건수 데이터는 평균에서 멀리 떨어진 값이 많아 log-scale로 시각화를 했다.

다음 그림의 1행의 데이터의 색깔은 고장 건수를 의미한다. 다음에 시각화 한 자료에서 위의 색깔을 사용할 것이기 때문에 대략적인 분포와 색깔을 표시했다.

2행은 따릉이 모델을 기준으로 색깔을 나는 것이다. 모델 간 운영 기간의 차이 때문에 사용 시간(x축)에 있어서 모델별 차이가 크다. 1행 2열을 보면 총 이동 거리가 3,269km(약 e^14) 전후인 자전거에서 잔고장이 많았다. 선형으로 비례하는 것은 아니지만, 이동 거리가 많은 자전거는 잔고장이 많아질 수 있다는 것을 알 수 있다.



1행의 그래프에서 데이터의 분포가 제대로 보이지 않아, 이를 log-scale 하여 2, 3행에 나타냈다. 2행은 고장 횟수로 색칠한 자료고, 3행은 자전거 모델(종류)에 따라 색칠했다.

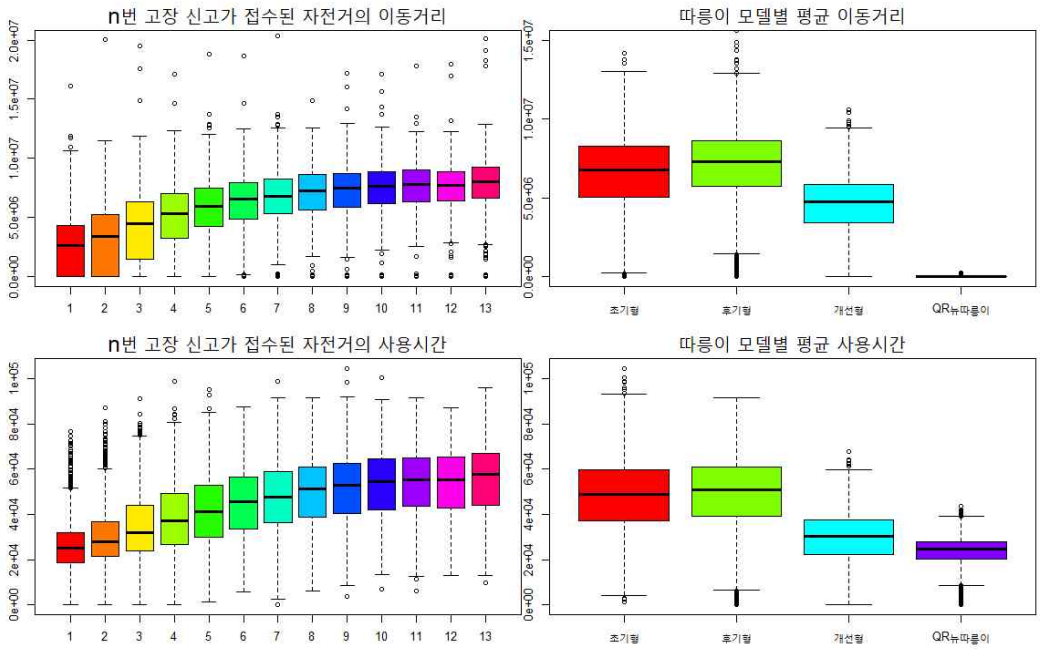


- x 축 : 자전거별 총 이동시간
- y 축 : log(총 이동 거리 + e^2)

위 그림 2행의 데이터를 보면 원점과 가까울수록 붉은 계열, 원점과 멀수록 푸른 계열을 띄는 것을 알 수 있다. 이는 자전거의 사용 시간과 이동 거리가 클수록 고장이 잘 난다는 것을 의미한다. box plot과 ANOVA 검사를 사용하여 분석해본다.

3.3. 분산분석

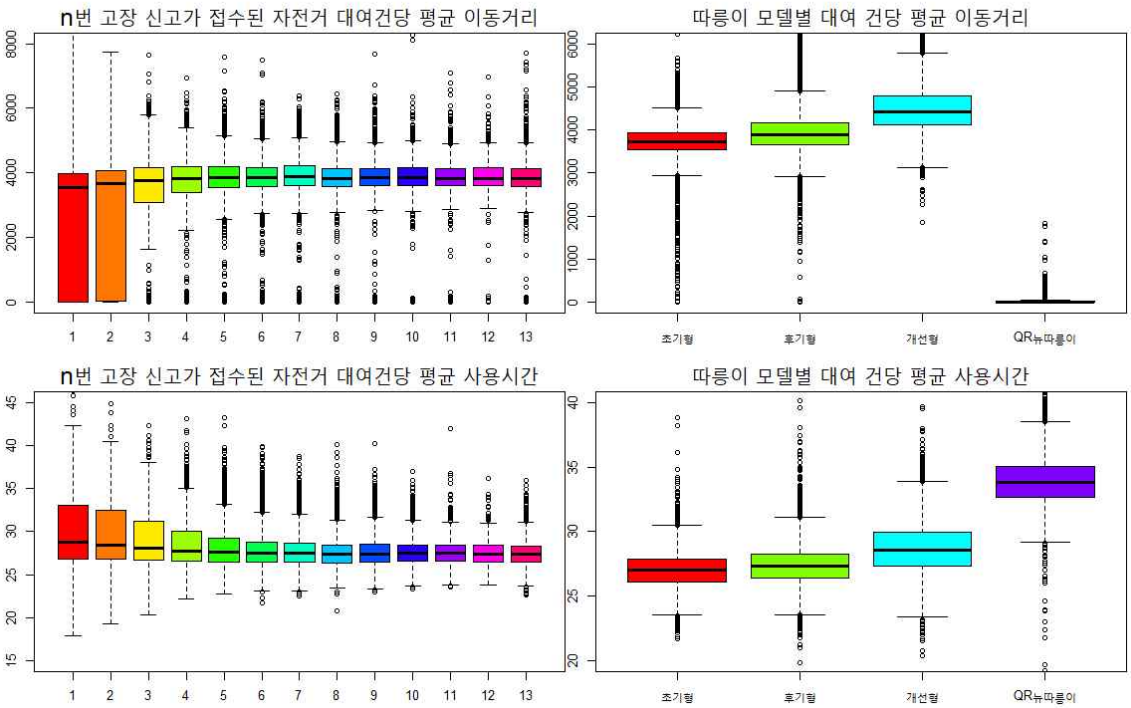
분산분석을 돌리기 전에 '자전거 별 운영기간'에 의한 차이를 고려할지에 대해 생각을 해봤다. 잠깐 고민을 해보니 두 상황에 대해서 각각 분석을 해보려고 한다. 우선 사용기간의 영향을 무시하지 않고 데이터를 살펴본다. 아래 그림은 자전거의 고유번호(ID)를 key값을 이동거리와 사용시간을 모두 '합계'한 자료다.



고장 횟수에 따른 사용시간 : F = 918.9 , P <= 2e-16 ***
고장 횟수에 따른 이동거리 : F = 355.0 , P <= 2e-16 ***
자전거 모델에 따른 사용시간 : F = 5244 , P <= 2e-16 ***
자전거 모델에 따른 이동거리 : F = 3546 , P <= 2e-16 ***

많이 고장 나는 자전거는 평균적으로 사용 시간과 이동 거리가 큰 '오래된 자전거'라는 것을 알 수 있다.

이번에는 운영 기간에 의한 영향을 배제하고 살펴보려고 한다. 이동 거리와 사용 시간을 대여 건수에 대해 평균을 내면, 운영 기간에 의한 영향이 보정되리라 생각했다.



고장 횟수에 따른 (대여 건당 평균) 사용시간 : F = 146 , P <= 2e-16 ***
고장 횟수에 따른 (대여 건당 평균) 이동거리 : F = 57.53 , P <= 2e-16 ***
자전거 모델에 따른(대여 건당 평균) 사용시간 : F = 9578 , P <= 2e-16 ***
자전거 모델에 따른(대여 건당 평균) 이동거리 : F = 2643 , P <= 2e-16 ***

QR 뉴 따릉이 데이터의 이동 거리 데이터가 유실되어 있음을 그것을 알게 됐다. 1행 2열의 표를 보면 다수의 뉴 따릉이의 이동 거리가 0으로 잡혀있다. 뉴 따릉이의 모델은 GPS 센서가 없나보다. 출발 대여소와 도착 대여소의 위도와 경도, 그리고 다른 자전거의 데이터를 사용하여 결측값을 채우는 모델을 만들어 데이터를 보정할 수 있을 것 같은데, 시간관계상 생략한다.

뉴 따릉이의 데이터 이상치의 영향으로 1행 1열 "1, 2, 3 번 고장' 테이터(뉴 따릉이의 지분이 큼)의 이동 거리 평균과 분산이 낮게 측정된다. 뉴 따릉이의 데이터의 시간은 웹서버를 통해 관리되고 있어 2행 2열에서 이상치는 발생하지 않는다.

따릉이 모델 중 '개선형'과 '뉴 따릉이'의 이용 시간 평균이 사용 시간이 높게 나오는데, 2017년 1월부터 시행된 따릉이 2시간 이용권의 영향으로 생각된다. 그전에는 정기권 1시간 이내 사용만 가능했다. 2행 2열 뉴 따릉이의 사용 시간 평균이 높은 것이 2행 1열 '1, 2, 3번 고장' 데이터의 분산으로 나타난다.

추후 연구

1. 생존 분석 (캐플런 마이어) 모델 공부하여 따릉이 데이터의 예상 수명을 구할 예정이다.
현재 고장과 수리가 반복되는 모형에 대해서 처리하는 방법을 찾아보고 있다.

2. 뉴 따릉이의 이동 거리를 예측하는 모형을 만들어 결측값을 보정 해볼 것이다.
공공 데이터 따릉이 대여소 목록을 보면 '위도, 경도, 지역구'에 관한 자료가 나와 있다.

###
이 문서에서 사용한 코드 및 자세한 내용은 첨부파일 및 깃허브에 게시했습니다.
깃허브 주소 : https://github.com/ParkWonBin/R_Seoul_Bike_DataAnalisys