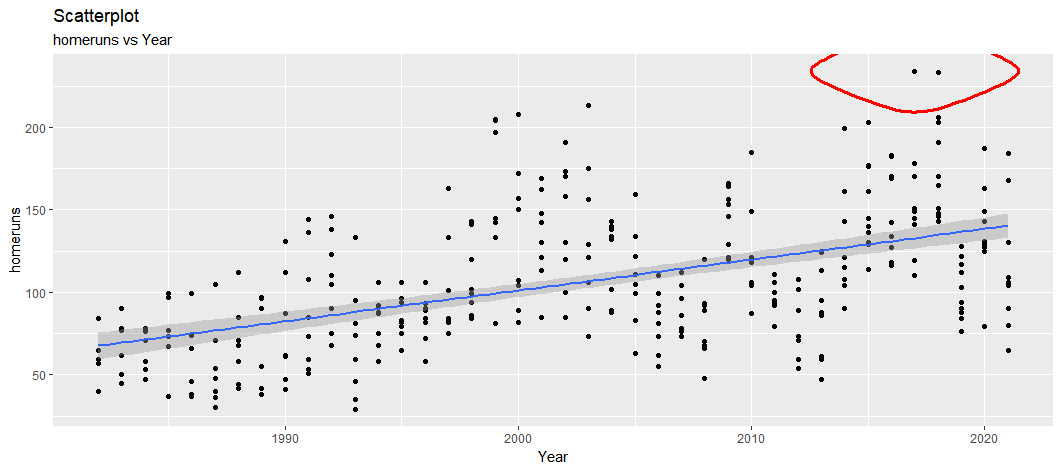
**<데이터셋 설명>**

이번 과제를 위해 선택한 데이터는 kaggle에서 찾은 Baseball KBO Batting Data(1982-2021)입니다. 데이터 세트에는 1982년부터 2021년까지 KBO 시즌별 연간 정규시즌 타격 데이터가 담겨 있습니다. 주어진 정보가 년도,팀이름,경기당 득점,평균연령,경기수,타석,안타수,2루타수,3루타수,홈런,타점,도루,볼넷,삼진,타율,도루 외에도 여러 야구 지표들이 있지만 앞에서 언급한 지표들 위주로 그래프를 작성하였습니다.

**<데이터셋 선정 이유>**

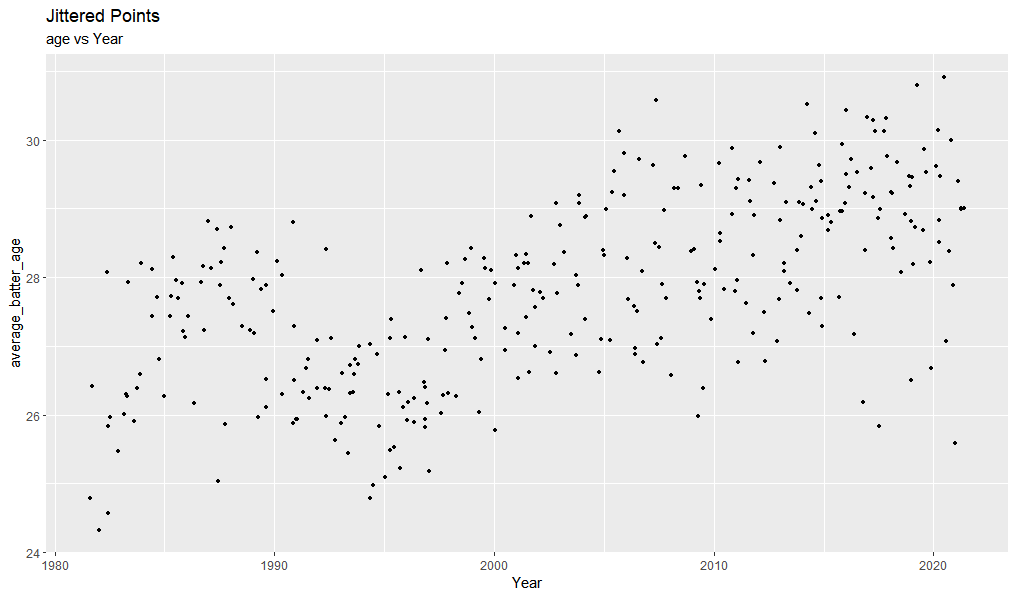
최근 포스트 시즌이 진행되고 있기도 하고 제가 어릴 적부터 응원하던 팀인 lg 트윈스가 가을야구에 진출해서 야구와 관련된 데이터를 분석해보고 싶었습니다. 전체적인 정보의 그래프 뿐만 아니라 lg 트윈스와 관련된 데이터를 통해 다양한 그래프를 그려 보았습니다. 나아가 야구와 같은 스포츠 분야에서 데이터 분석의 중요성이 최근 강조되고 있어서 부족하지만 관련한 주제를 선정하고 다양한 그래프를 그려보고자 하였습니다.

**1 Scatter plot with Encircling**



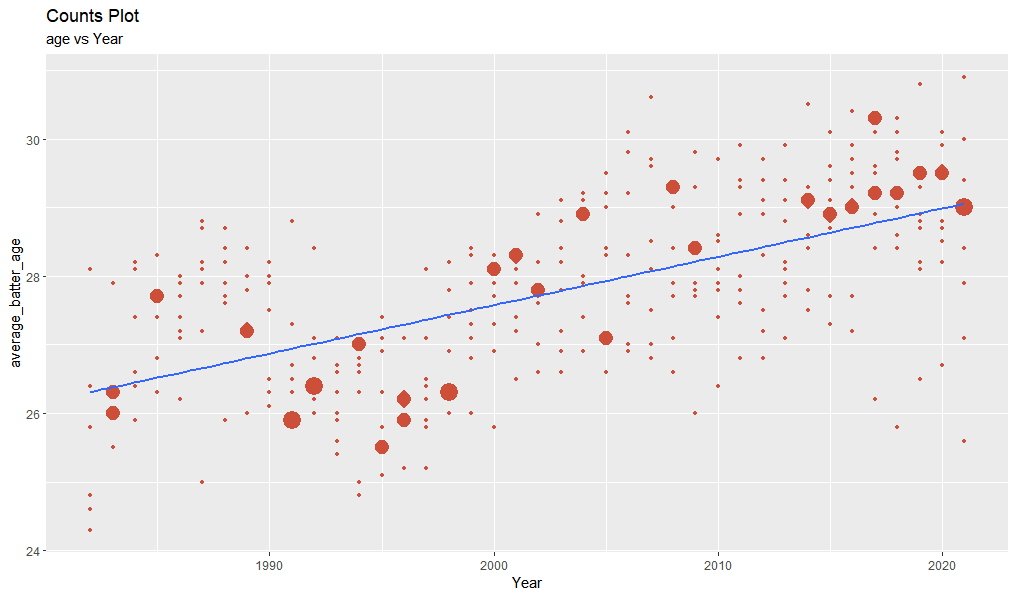
x축을 년도 y축을 홈런 개수로 하는 산점도 입니다. 산점도를 통해 두 변수 년도와 홈런 개수 사이의 관계에 대해 알 수 있습니다. 년도가 증가할수록 홈런의 개수가 증가한다는 것을 알 수 있습니다. 이 그래프에서 가장 홈런의 많이 친 팀과 년도 그리고 당시 총 홈런 개수를 표현해주기 위해 ggalt 패키지의 gem\_encirlce()의 사용하였습니다. 그 결과 빨간색 색상과 3의 크기인 두께로 그래프 상에서 동그라미 쳐진 것을 확인할 수 있습니다. 당시 팀은 2017년도 SK로 홈런 총 개수는 234개입니다.

**2. Jittered**



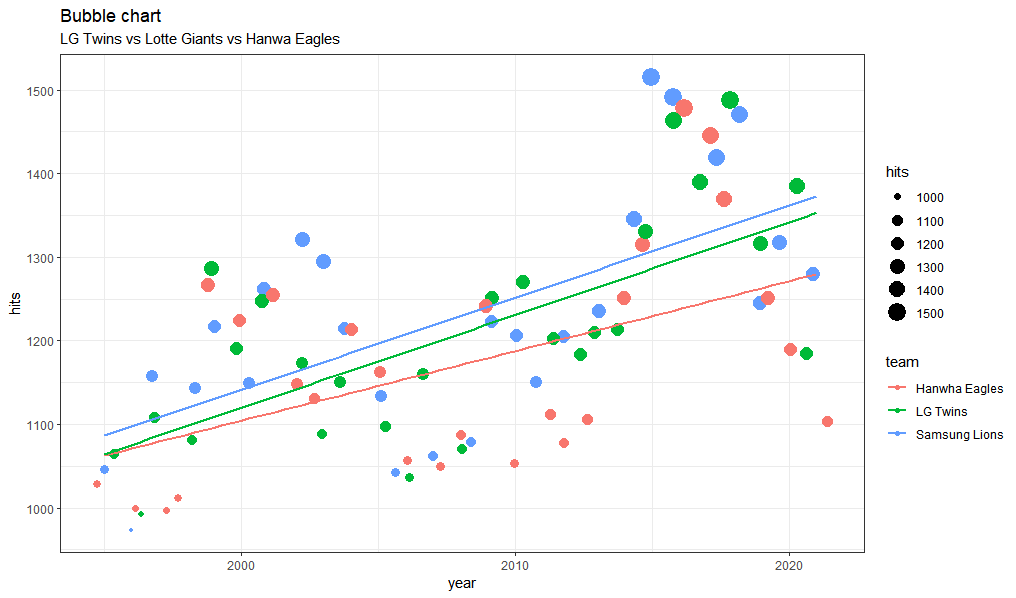
x축을 년도 y축을 각 팀 선수들의 평균 나이로 하는 산점도를 그리면 중복되는 값들 때문에 총 개수 보다 적은 수의 점들이 표현된다. 이를 해결하기 위해 jitter를 사용하여 겹치는 점들은 너비 인수에 의해 무작위로 점등되는 형태로 그립니다. 그래프를 보면 전체적으로 팀들의 평균 연령이 증가하는 경향은 있지만 의미 있는 특징을 찾기는 어려워 보입니다.

**3. Counts plot**

. 

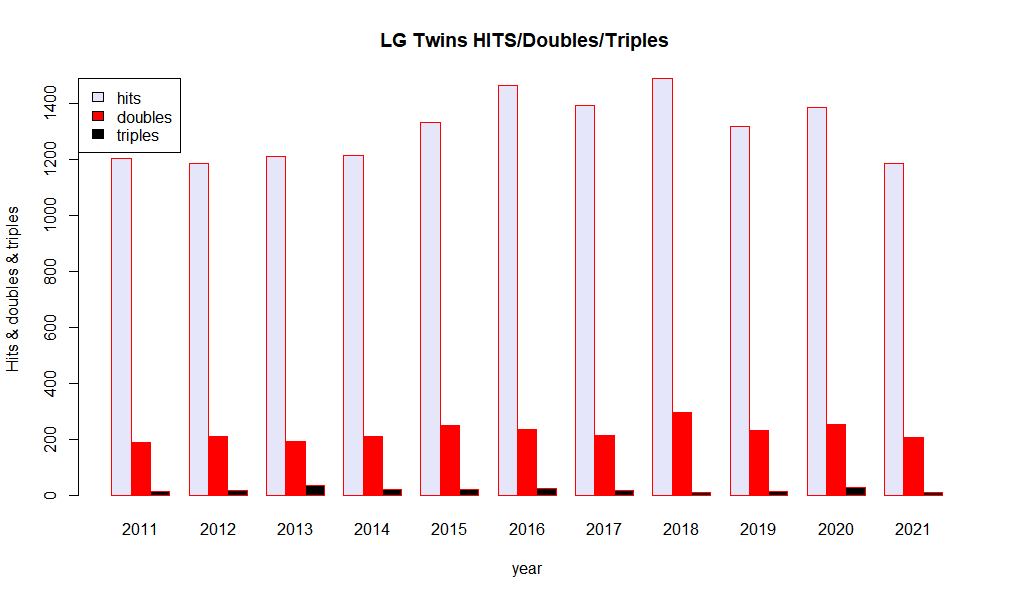
2번 문제와 동일한 변수로 x축은 년도 y축은 평균 연령입니다. 여기서도 데이터의 점들이 겹치는 문제가 발생하는데 이를 해결하기 위하여 카운트 plot을 이용하였습니다. 점이 겹칠수록 원의 크기가 커지게 그려집니다. 그래프에서 마찬가지로 전체적으로 평균연령이 증가함을 알 수 있고 파란색 직선 주위에 중복되는 값들(큰 원)이 많이 있습니다.

**4.Bubble plot**



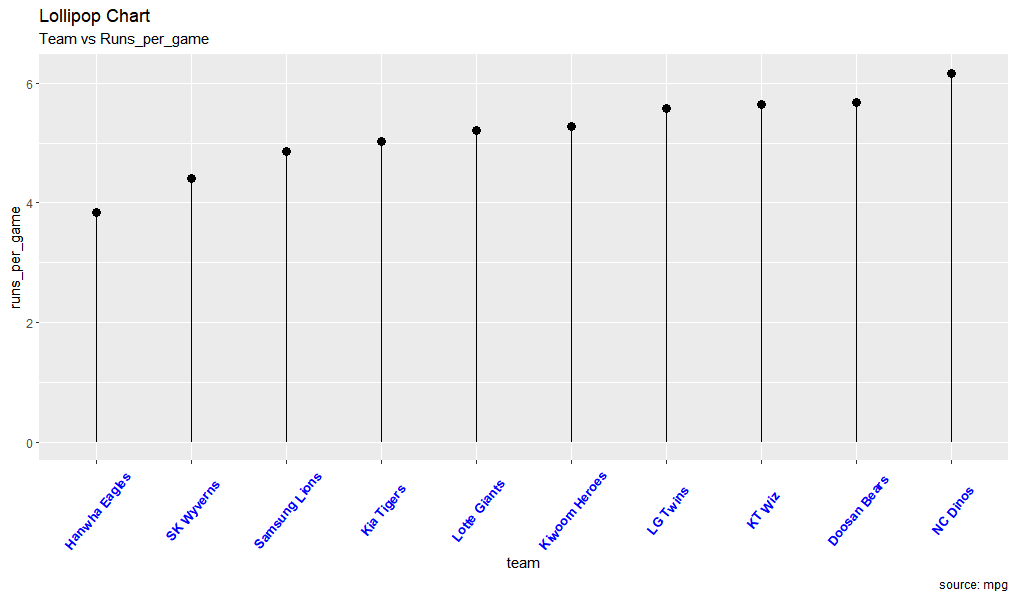
산점도에서는 최대 2개의 변수 사이의 관계 밖에 표현이 불가능했습니다. 이 한계를 극복하기 위해 버블 plot을 그려보았습니다. 먼저 제가 좋아하는 lg그리고 최근은 명문구단에서 벗어났지만 2000년대 항상 강팀으로 있었던 삼성 그리고 항상 하위권에 있는 환화 3개의 팀을 선정하였습니다. 년도는 빙글레 이글스가 환화 이글스로 이름이 바뀐 1994년 이상부터 선택하였습니다. x축의 년도 y축의 안타 개수로 하여 그래프를 그렸습니다. 그래프를 보면 어느 정도 예상한 결과를 알 수 있습니다. 년도에 따른 안타의 개수의 경향성을 보여주는 직선을 보면 강팀인 삼성과 약팀인 환화 사이의 차이를 알 수 있으며 이는 안타 개수와 팀의 순위 사이의 관계성을 보여준다고 생각합니다.

**5. Bar plot**



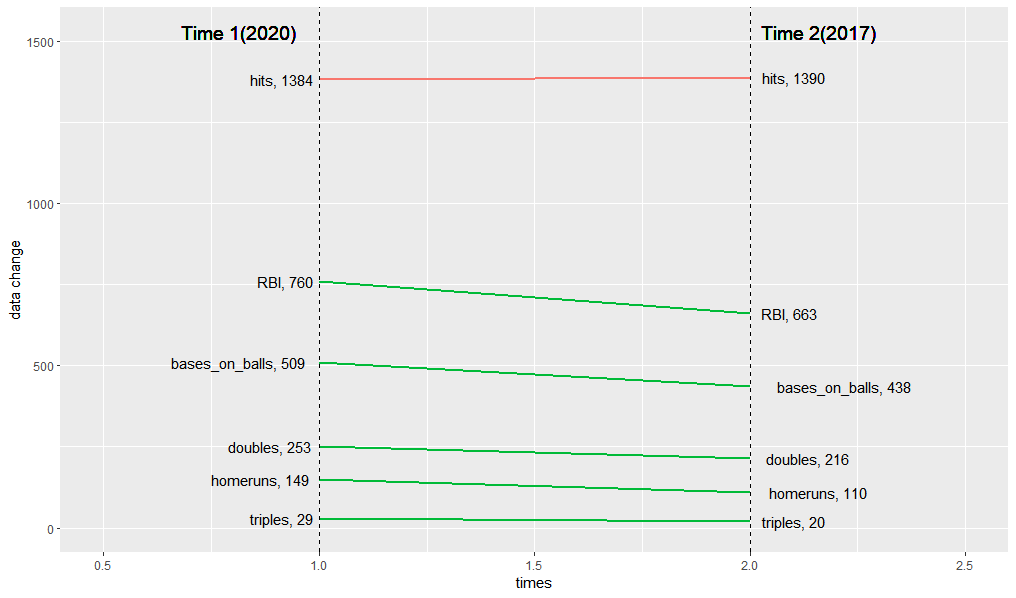
가장 좋아하는 팀인 lg와 야구를 보기 시작하던 2011년부터의 lg의 안타,2루타,3루타 개수를 bar plot을 이용해 그려보았습니다. 각 바의 테두리를 lg를 상징하는 빨간색으로 표현하였고 바는 안타는 라벤더 2루타는 빨간 3루타는 검정색으로 채워 lg를 대표하는 색상들로 꾸며 보았습니다.

**6. Lollipop Chart**

****

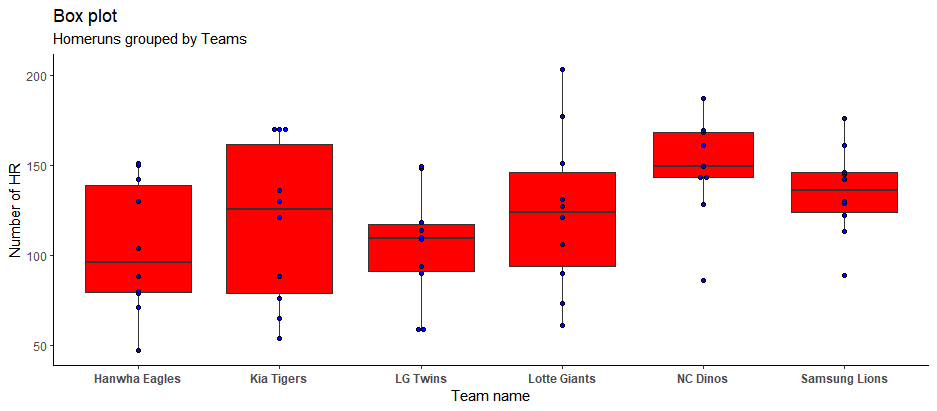
기본적으로 위에서 그린 bar plot과 동일한 형태라고 할 수 있는데 조금 더 가독성 있는 그래프입니다. 2020년 각 팀별의 승률에 대한 차트로 데이터 프레임을 경기당 득점이 낮은 순서대로 정렬한 다음 그래프를 그렸습니다. 일반적으로 경기당 득점이 높을수록 순위가 높을 것이라고 추측 할 수 있습니다. 실제로 2021년 팀들의 순위와 비교해보면 1위였던 NC의 경기당 득점이 가장 높고 꼴등인 환화의 득점이 가장 낮음을 파악할 수 있었습니다.

**7. Slope chart**

****

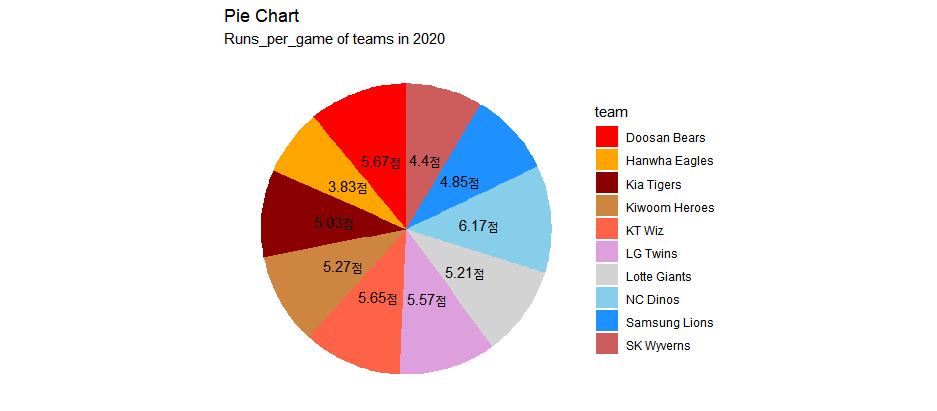
기울기 차트를 이용해 특정 두 시점에서 변화의 차이를 표현하는 그래프를 그려 보았습니다. 선택한 데이터는 lg의 안타,타점,볼넷,2루타,홈런,3루타 의 2017년과 2020년의 값입니다. 2017은 lg 감독 류중일의 부임 첫해이고 2020년은 마지막 해였습니다. 이 그래프를 통해 그가 감독으로 있는 동안 공격적인 측면의 지표 변화에 대해 분석해 볼 수 있었습니다. 안타를 제외하고 17년도에 비해 20년도의 모든 값이 증가하였고 이는 초록색 직선의 형태로 표현됩니다. 이를 통해 류중일 감독은 lg의 공격성 증가에 큰 기여를 하였다고 해석할 수 있었습니다.

**8. Box plot (with dots)**

****

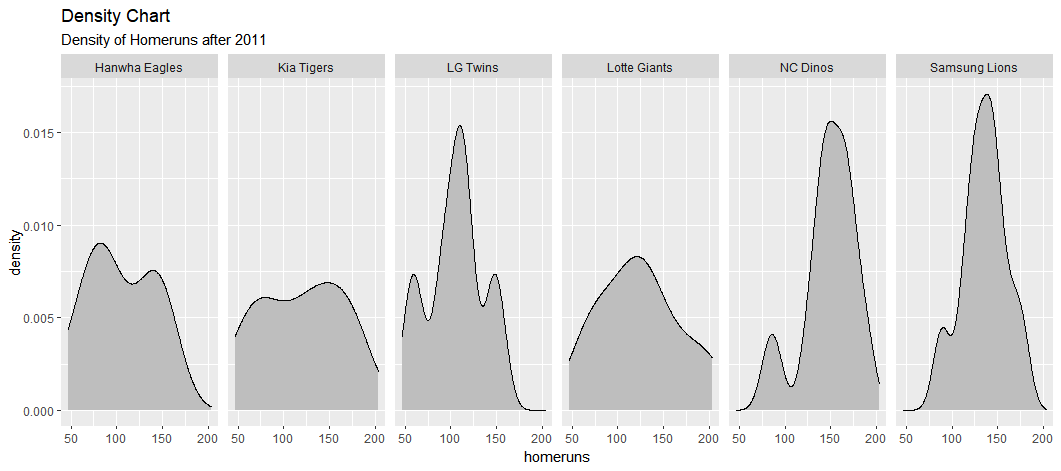
Box plot은 특정 데이터의 중간값 범위 및 특이치 등 다양한 요소들의 표현하고 있기 때문에 분포를 분석하는데 용이합니다. 먼저 2011년 이후 위의 6팀(환화,기아,lg,롯데,NC,삼성)들의 홈런 분포에 관해 그래프를 그려 보았습니다. 조금 특이한 팀이 롯데인데 박스의 크기에 비해 박스 밖의 직선이 상대적으로 매우 큰 것을 확인 할 수 있습니다. 이는 년도별로 홈런 개수의 편차가 크다고 해석할 수 있습니다. 반면 NC 다이노스는 박스의 높이가 가장 작기 때문에 반대로 편차가 작음을 알 수 있습니다. 기존의 box plot에다가 각 데이터 값들의 파란색 점으로 찍어 주어 가독성의 높여보았습니다. 실제로 nc는 점들이 중간값 주위로 많이 모여서 분포하고 있습니다. 박스 안의 진한 검은색 줄이 중간 값을 표현하는데 이를 통해 각 팀들의 평균적인 홈런 능력에 관해 추정해볼 수 있었습니다. 가장 낮은 환화는 홈런타자가 적은 팀 반면 삼성과 nc는 높은 수치를 가지는 홈런 타자가 많은 팀이라는 결과를 얻을 수 있습니다.

**9. Pie chart**

****

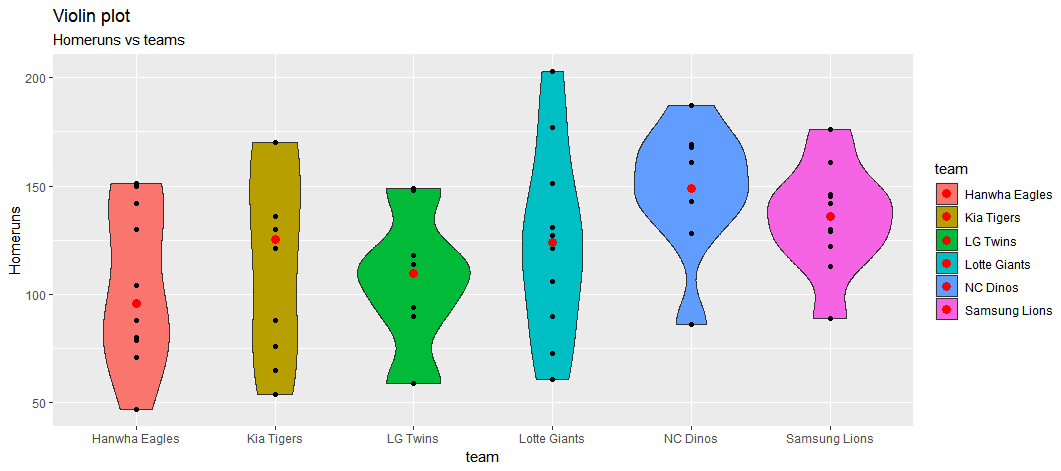
2020년도 각 팀별 득점수를 pie chart로 그려보았습니다. 색상은 각 팀의 대표색으로 바꾸어 표현하였고 파이차트 내부에 각 값들의 출력해보았습니다. 그래프 분석 결과 NC 다이노스의 팀 득점이 평균 6.17점으로 가장 크고 환화 이글스의 팀 득점이 평균 3.83점으로 가장 낮은 것을 확인 할 수 있었습니다.

**10.Density chart**



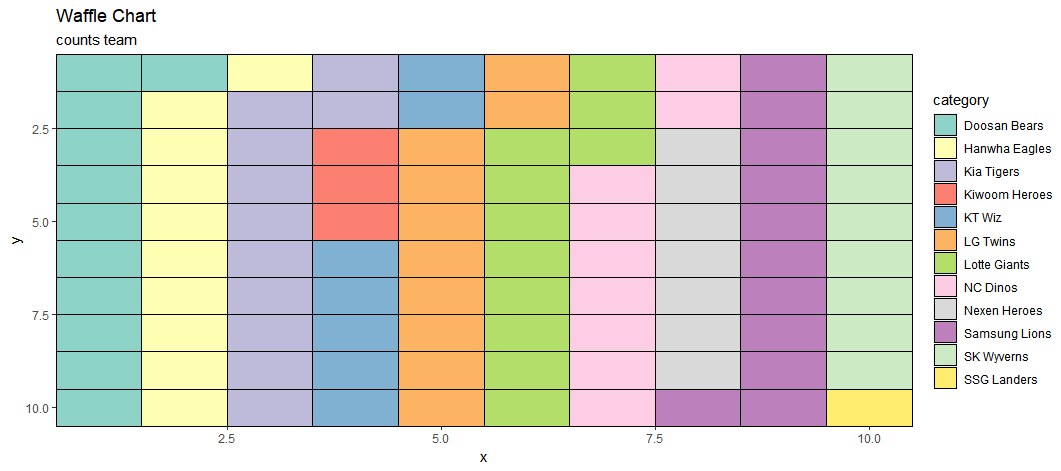
선택한 데이터와 분석하고자 하는 내용은 8번과 동일합니다. 환화, 기아, lg ,롯데, NC, 삼성 6개 팀의 2011년 이후의 홈런 개수를 밀도 그래프를 이용하여 그려보았습니다. 밀도 그래프의 특징은 회색으로 색칠된 면적을 합이 1이라는 점입니다. facet\_grid를 통해 팀별로 정렬하였습니다. 8번의 boxplot보다 density chart로 표현했을 때 한눈에 더 잘 들어오는 가독성을 가지는 측면이 있습니다. 결과를 보아도 확실히 강팀인 NC,삼성의 홈런 개수분포가 높게 치솟아 있는 모습을 확인 할 수 있습니다.

**11. Violin plot**



선택한 데이터와 분석하고자 하는 내용은 8번과 동일합니다. 8번의 boxplot 에서는 평균/중앙값 같은 요약 통계량이 표시되는 반면, 바이올린 그림에서는 데이터의 전체 분포가 표시됩니다. 그래프에서 중앙값들의 빨간색 점으로 표현하였습니다. 8,10번과 동일한 데이터를 표현한 그래프인데 팀별 홈런 분포의 값을 분석하는데 있어서 violine plot이 가장 효과적이라고 할 수 있습니다. 실제로 그래프 상에서 보면 nc와 삼성의 바이올린이 위쪽으로 그려진 것을 확인 할 수 있습니다.

**12. Waffle chart**



전체 데이터에서 최근 100개의 행을 선택해서 팀의 개수에 관한 데이터로 waffle chart를 그려보았습니다. 상대적으로 많은 칸의 차지하는 롯데, 삼성,기아는 역사가 오래된 팀인 반면 최근 이름이 변경됨 키움과 ssg landers는 칸의 개수가 적다는 점을 확인 할 수 있었습니다.