**Little Café**

Hyunjae Lee / Hosoo Lee / Duyoung Choi

**#1 How many unique observations to you have?**

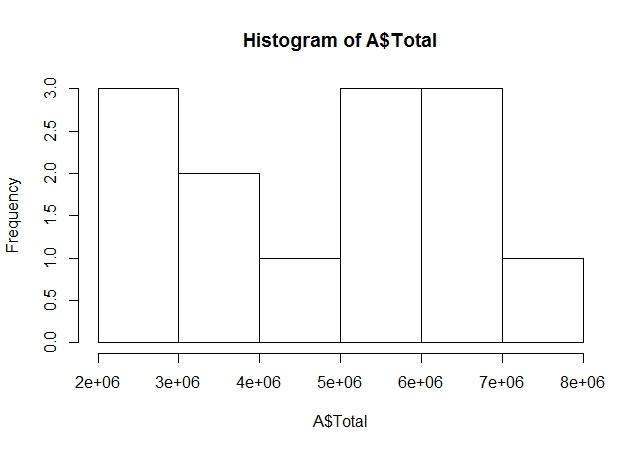
We compare not just ranking and the number of spectator, but also main factors that arouse spectators’ interest such as batting average and earned run average.

**#2 What information / features / characteristics do you have for each observation?**

Based on the KBO’s record, for estimating the correlation between team performance and the number of spectator, we collect data about attendance, ranking, winning rate, batting rate and earned run average from 2002. Now, Korean professional baseball operates 10 clubs, but KT(join to first division in 2015) and NC(join to first division in 2013) are excluded because they have less samples. Nexen is also joined to first division in 2008, so have comparatively less samples. But it was formerly Hyundai that was joined to first division until 2007, so for convenience, we combine two teams’ records.

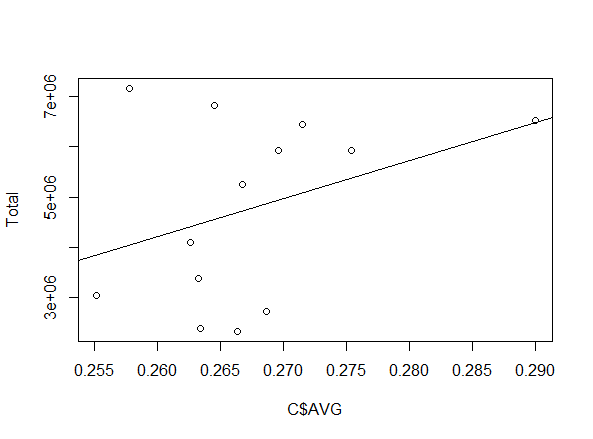
**#3 What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features (show a histogram)?**

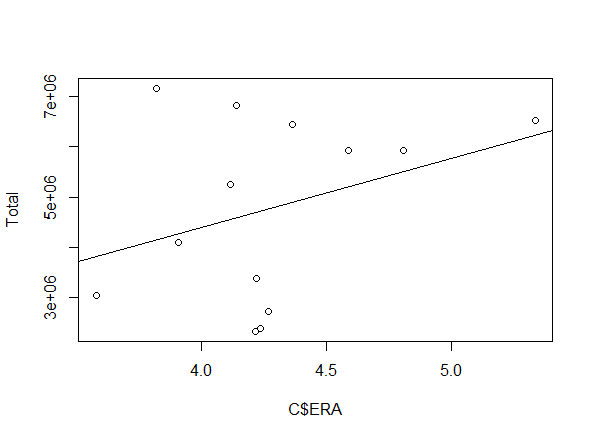
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AVG | ERA | RANK | TOTAL |
| Min. | 0.2430 | 3.220 | 1.00 | 2331978 |
| Median | 0.2680 | 4.220 | 4.50 | 5256332 |
| Mean | 0.2673 | 4.277 | 4.50 | 4770013 |
| Max. | 0.3010 | 6.350 | 8.00 | 7156157 |
| SD | 0.01264825 | 0.6484333 | 2.302384 | 1819221 |



We select attendance(total) for core features. Our purpose is analyzing the correlation between performance and attendance. So ranking, batting rate and earned run average are used to sub data, and performance is selected to core function.

**#4 Are there obvious trends in the data (over time, across subgroups, etc.), and are the differences statistically significant?**

****

****

Just whole attendance is in the increasing tendency and attendance per team also shows the entirely increasing tendency. Correlation between batting rate, earned run average, ranking and attendance is generally little and total is also the same.

**#5 What are the other salient aspects of the data (e.g. geospatial factors, text, content, etc.)**

Our data is limited in Korean professional baseball and first division. But we selected ranking, batting average, ERA(earned run average) as our sub data, which are the most representative record on team total, offense and defense. High ranking does not always means high batting average and low ERA. So, we tried to think of a variety of ways to do the analysis.

**#6 Provide a bullet-list of the next 5-10 tasks you will perform in analyzing your dataset.**

* Correlation analysis between team and total league attendance
* Correlation analysis between team performance and team attendance
* Checking whether the null hypothesis is true or not by t-test
* Analysis and interpretation of t-test result
* Supplementation from the interpretation result

**Little Cafe**



Hyunjae Lee / Hosoo Lee / Duyoung Choi