

Spring 2022

Code ▾

# GE 461 Introduction to Data Science

Statistical Models by Savaş Dayanık

- Prerequisites
- March 1: Exploratory data analysis
- Next time: A linear regression model
- Project (will be graded)
- Bibliography

## Week 5: Advertising and Promotion

The Dodgers is a professional baseball team and plays in the Major Baseball League. The team owns a 56,000-seat stadium and is interested in increasing the attendance of their fans during home games. *At the moment the team management would like to know if bobblehead promotions increase the attendance of the team's fans?* This is a case study based on Miller (2014, chap. 2).

Code



Figure 1: 56,000-seat Dodgers stadium (left), shirts and caps (middle), bobblehead (right)

The 2012 season data in the `events` table of SQLite database `data/dodgers.sqlite` contain for each of 81 home play the

- month,
- day,
- weekday,
- part of the day (day or night),
- attendance,
- opponent,
- temperature,
- whether cap or shirt or bobblehead promotions were run, and

- whether fireworks were present.

## Prerequisites

We will use R, RStudio, R Markdown for the next three weeks to fit statistical models to various data and analyze them. Read Wickham and Grolemund (2017) online

- Section 1.1 (<https://r4ds.had.co.nz/introduction.html#prerequisites>) for how to download and install R and RStudio,
- Chapter 27 (<https://r4ds.had.co.nz/r-markdown.html>) for how to use R Markdown to interact with R and conduct various predictive analyses.

All materials for the next three weeks will be available on Google drive ([https://drive.google.com/drive/folders/1ehZI2fF7awKOBik7YSuAbyGV7Sa2jOj\\_?usp=sharing](https://drive.google.com/drive/folders/1ehZI2fF7awKOBik7YSuAbyGV7Sa2jOj_?usp=sharing)).

## March 1: Exploratory data analysis

1. Connect to `data/dodgers.sqlite`. Read table `events` into a variable in R.
  - Read Baumer, Kaplan, and Horton (2017, chaps. 1, 4, 5, 15) (Second edition online (<https://mdsr-book.github.io/mdsr2e/>)) for getting data from and writing them to various SQL databases.
  - Because we do not want to hassle with user permissions, we will use SQLite for practice. I recommend PostgreSQL for real projects.
  - Open RStudio terminal, connect to database `dodgers.sqlite` with `sqlite3`. Explore it (there is only one table, `events`, at this time) with commands
    - `.help`
    - `.databases`
    - `.tables`
    - `.schema <table_name>`
    - `.headers on`
    - `.mode column`
    - `SELECT ...`
    - `.quit`
  - Databases are great to store and retrieve large data, especially, when they are indexed with respect to variables/columns along with we do search and match extensively.
  - R (likewise, Python) allows one to seemingly read from and write to databases. For fast analysis, keep data in a database, index tables for fast retrieval, use R or Python to fit models to data.

Code

Code

```
## # A tibble: 3 × 3
##   bobblehead fireworks      n
##   <fct>         <fct>    <int>
## 1 NO           NO        56
## 2 NO           YES        14
## 3 YES          NO        11
```

2. What are the number of plays on each week day and in each month of a year?

Table 1 and 2 summarize the number of games played on each weekday and month.

Code

Table 1: Number of games played in each weekday and month

month	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
APR	1	2	2	1	2	2	2
MAY	3	3	2	1	3	3	3
JUN	1	1	1	1	2	2	1
JUL	3	3	2	NA	1	1	2
AUG	2	2	3	1	3	2	2
SEP	1	1	1	1	2	3	3
OCT	1	1	1	NA	NA	NA	NA

Code

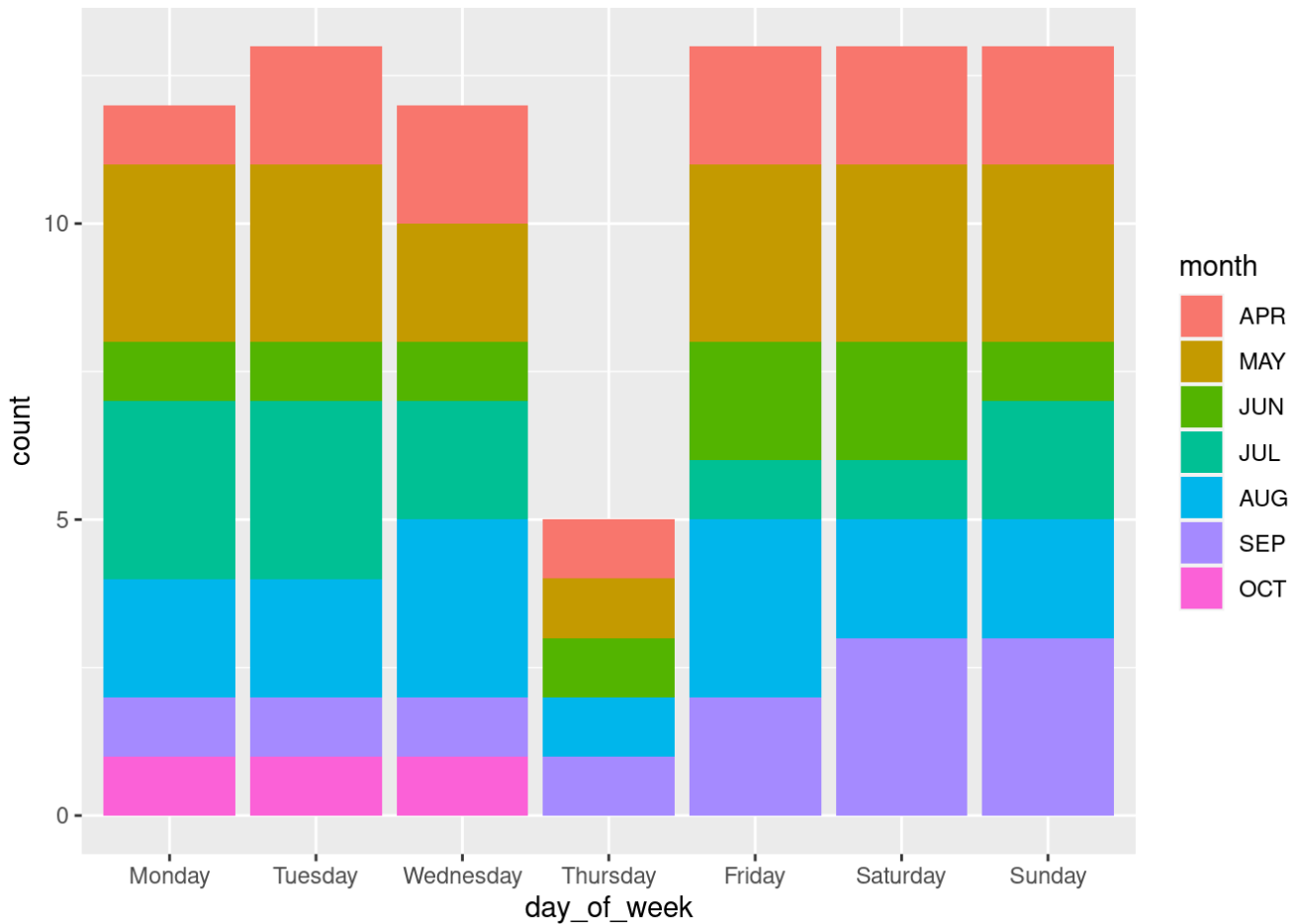


Figure 2: Barplot of counts of games for each weekday and month

Figure 3 shows your friend's (very good) suggestion of heatmap of total attendance versus weekday and month. The colors change from bright yellow to dark red as attendance increases. Default heatmap shuffles rows and columns so as to bring together weekdays and months with similar attendance. Here we see May, Aug, and Jul together within the months and Saturday, Friday, Sunday within the weekdays. Learn more about xtabs (cross-table) heatmap by typing `?xtabs` and `?heatmap` in the R console.

Code

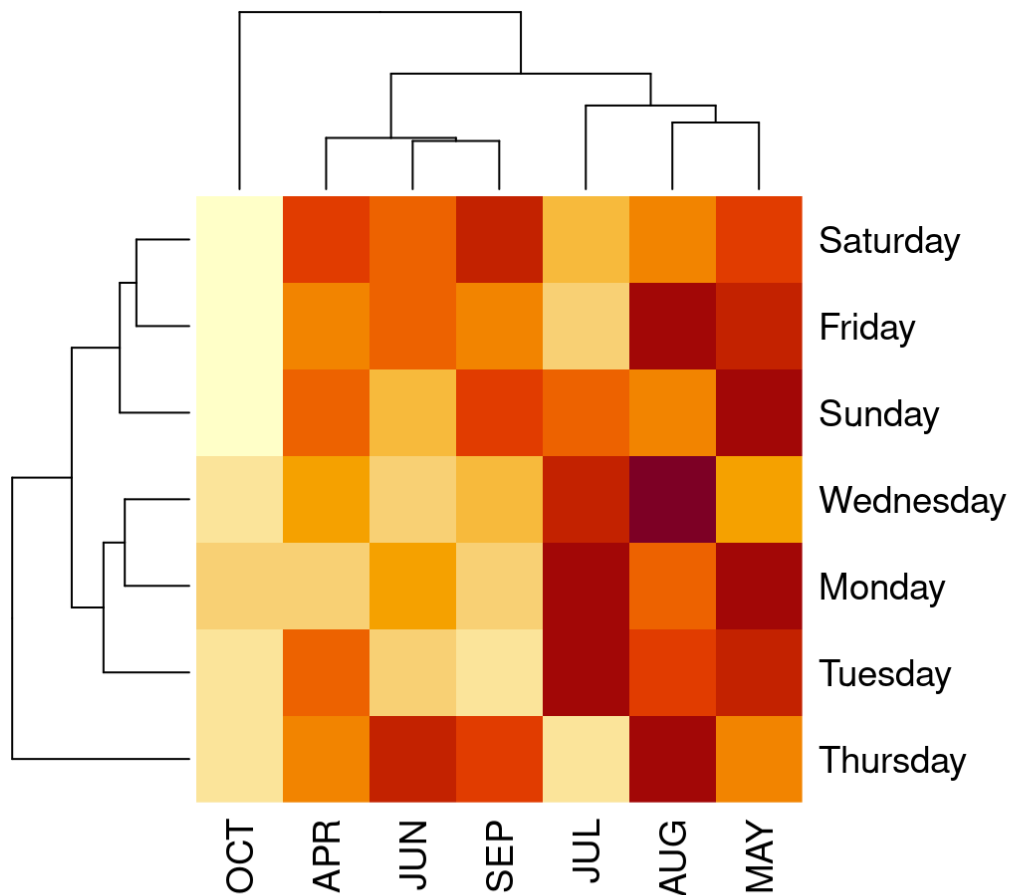


Figure 3: Heatmap of attendance versus weekday and month.

In Figure 4, I have added one more aes (colour) to capture day\_night information. To avoid overplotting, I replaced `geom_point()` with `geom_jitter()`. These plots were also illuminating. So let us thank your friend who suggested this one, too.

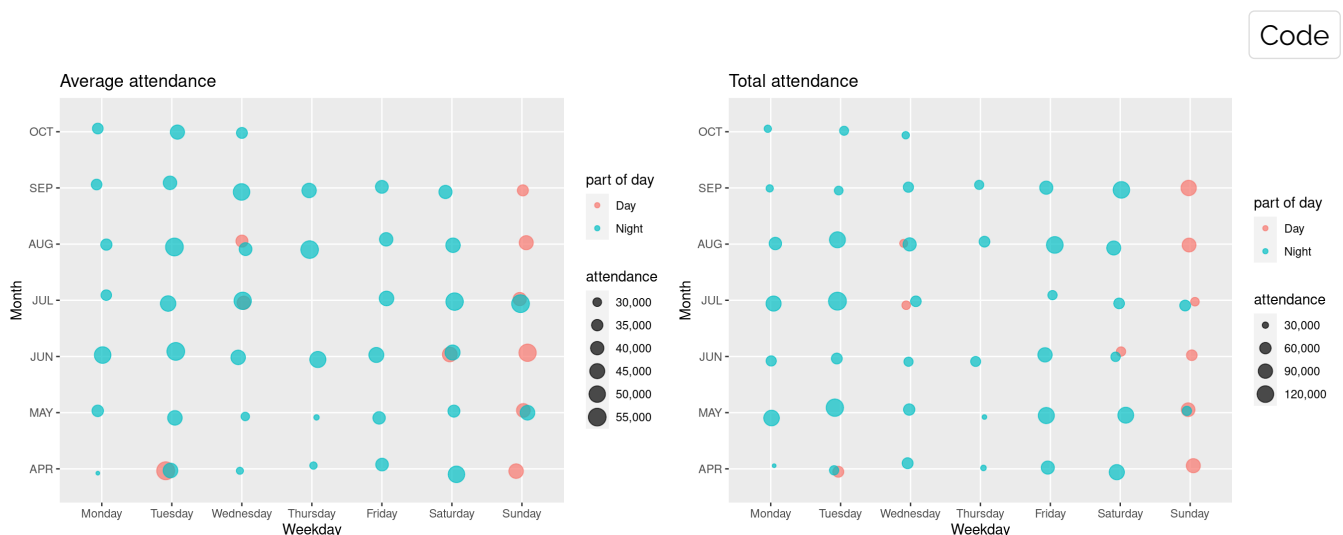


Figure 4: Average and total attendances on each weekday and month in each part of day

3. Check the orders of the levels of the `day_of_week` and `month` factors. If necessary, put them in the logical order.

Code

```
## [1] "Monday"    "Tuesday"    "Wednesday"  "Thursday"   "Friday"     "Saturday"
## [7] "Sunday"
```

Code

```
## [1] "APR" "MAY" "JUN" "JUL" "AUG" "SEP" "OCT"
```

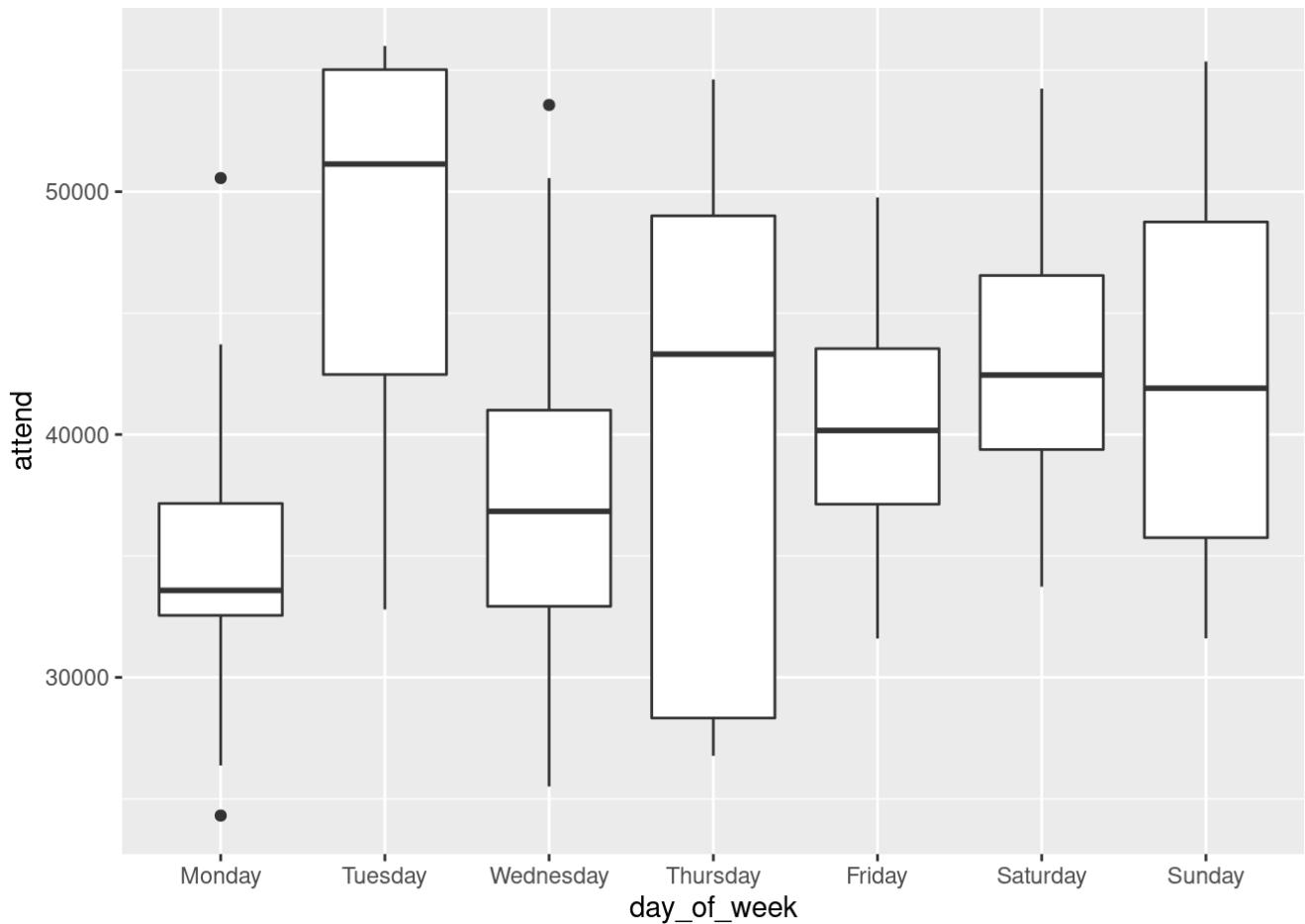
4. How many times were bobblehead promotions run on each week day?

Code

```
## # A tibble: 7 × 3
##   day_of_week Bobblehead Total
##   <fct>         <dbl> <dbl>
## 1 Monday           0     12
## 2 Tuesday          6     13
## 3 Wednesday        0     12
## 4 Thursday         2      5
## 5 Friday           0     13
## 6 Saturday         2     13
## 7 Sunday           1     13
```

5. How did the attendance vary across week days? Draw boxplots. On which day of week was the attendance the highest on average?

Code



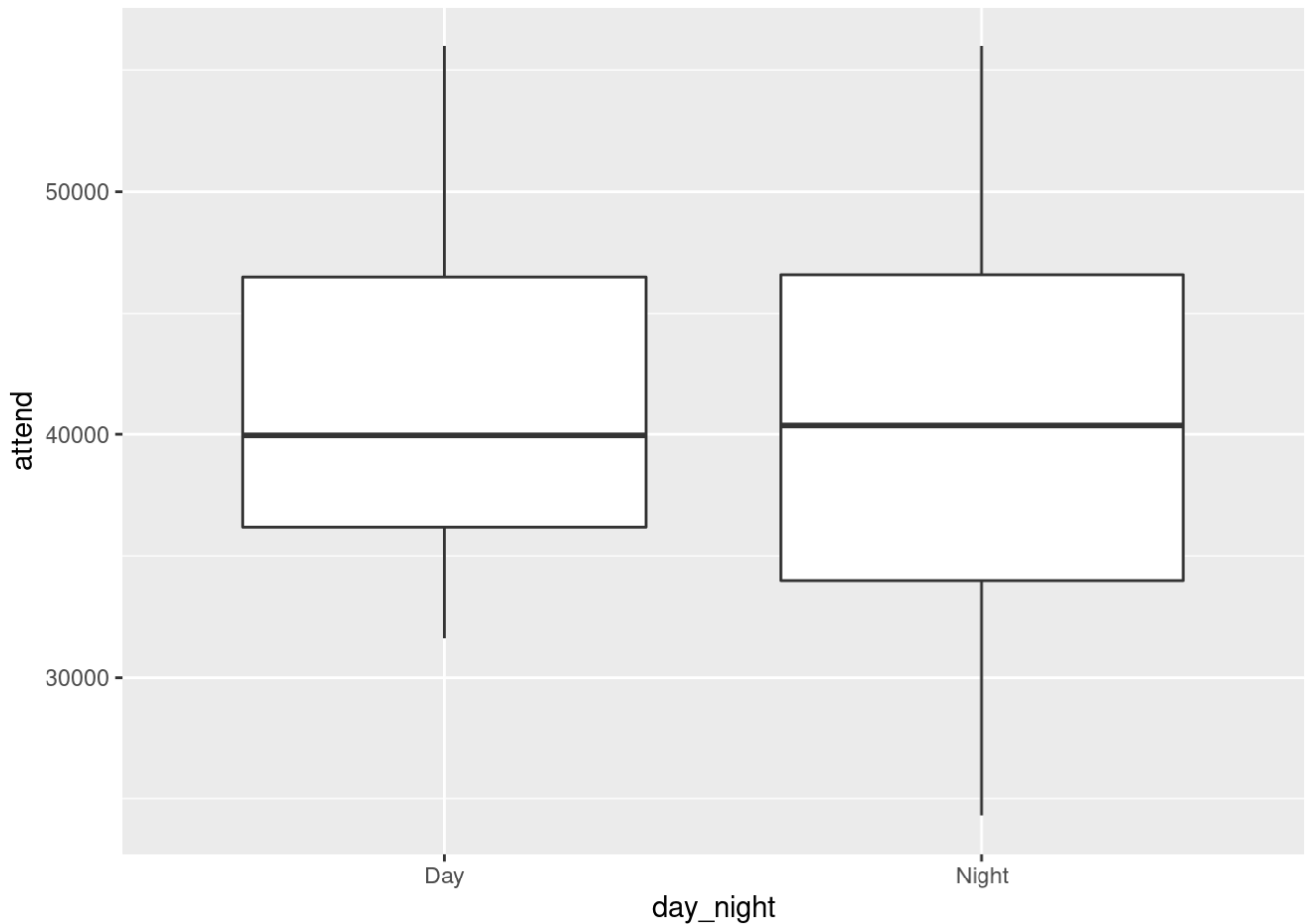
Code

```
## # A tibble: 5 × 12
##   month   day attend day_of_week opponent   temp skies day_night cap
##   <fct> <dbl>  <dbl> <fct>      <fct>    <dbl> <fct>  <fct>    <fc
##   <fct>
## 1 APR      10  56000 Tuesday   Pirates    19 Clear  Day      NO
## 2 AUG      21  56000 Tuesday   Giants     24 Clear  Night    NO
## 3 JUL       1  55359 Sunday    Mets       24 Clear  Night    NO
## 4 JUN      12  55279 Tuesday   Angels     19 Cloudy Night    NO
## 5 AUG       7  55024 Tuesday   Rockies    27 Clear  Night    NO
## # ... with 2 more variables: fireworks <fct>, bobblehead <fct>
```

6. Is there an association between attendance and

- whether the game is played in day light or night?
- Between attendance and whether skies are clear or cloudy?

Code

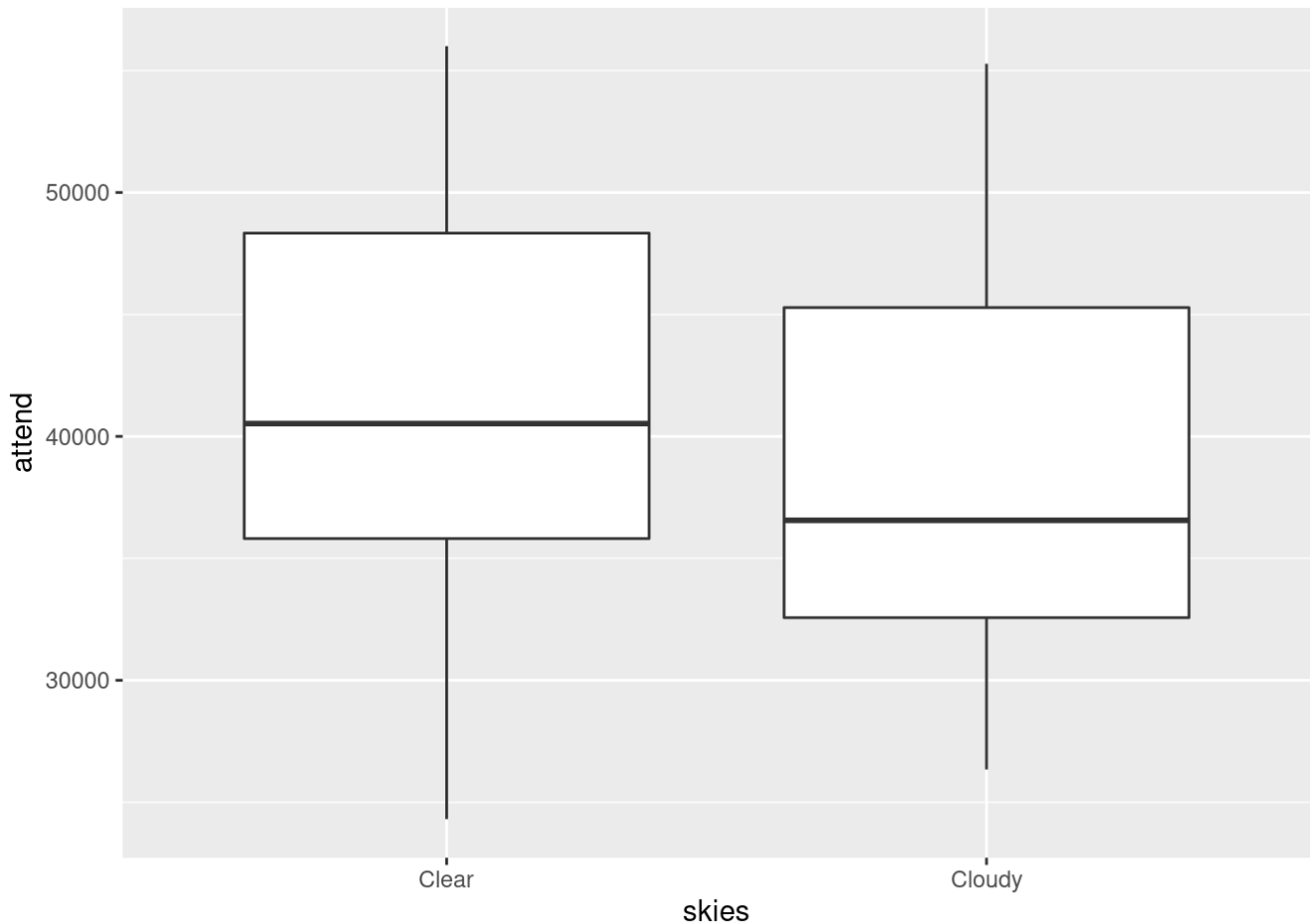
[Code](#)

```
##  
## Welch Two Sample t-test  
##  
## data: events$attend[events$day_night == "Day"] and events$attend[ev  
ents$day_night == "Night"]  
## t = 0.42851, df = 23.62, p-value = 0.6722  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3531.652 5380.397  
## sample estimates:  
## mean of x mean of y  
## 41793.27 40868.89
```

Since p-value (0.67) is large (greater than 0.05), we cannot reject null, which means there is no statistical difference between average attendance of games played in day and night.

[Code](#)





Code

```
##
## Welch Two Sample t-test
##
## data: events$attend[events$skies == "Clear"] and events$attend[events$skies == "Cloudy"]
## t = 1.2868, df = 27.664, p-value = 0.2088
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1741.315 7617.103
## sample estimates:
## mean of x mean of y
## 41729.21 38791.32
```

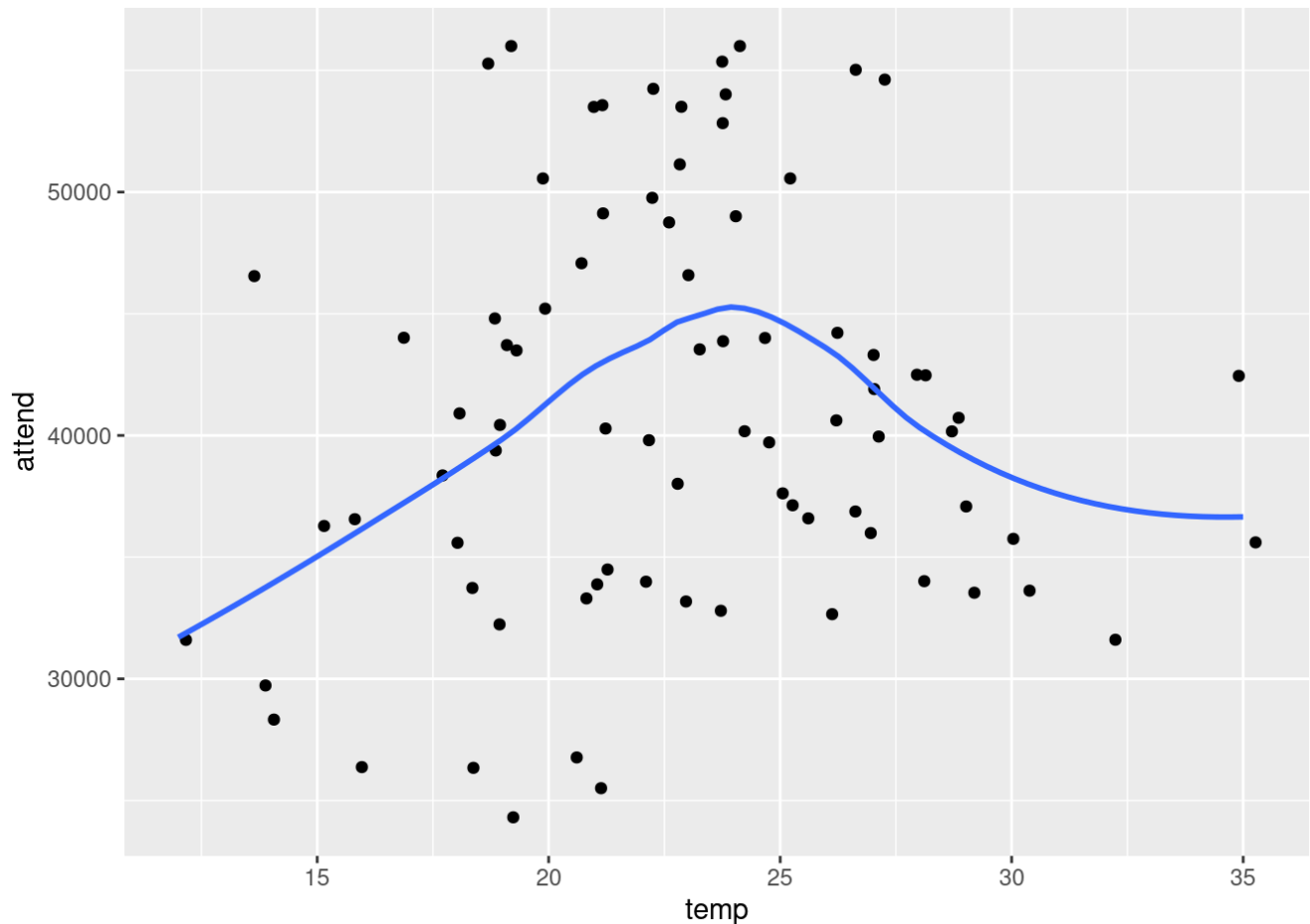
We do not see a statistically significant difference between the average attendance of the games played under clear and cloudy skies.

7. Is there an association between attendance and temperature?

- If yes, is there a positive or negative association?
- Do the associations differ on clear and cloudy days or day or night times?

Code

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



$$attend = \beta_0 + \beta_1 temp + \beta_2(temp - 23)^+ + \varepsilon_i$$

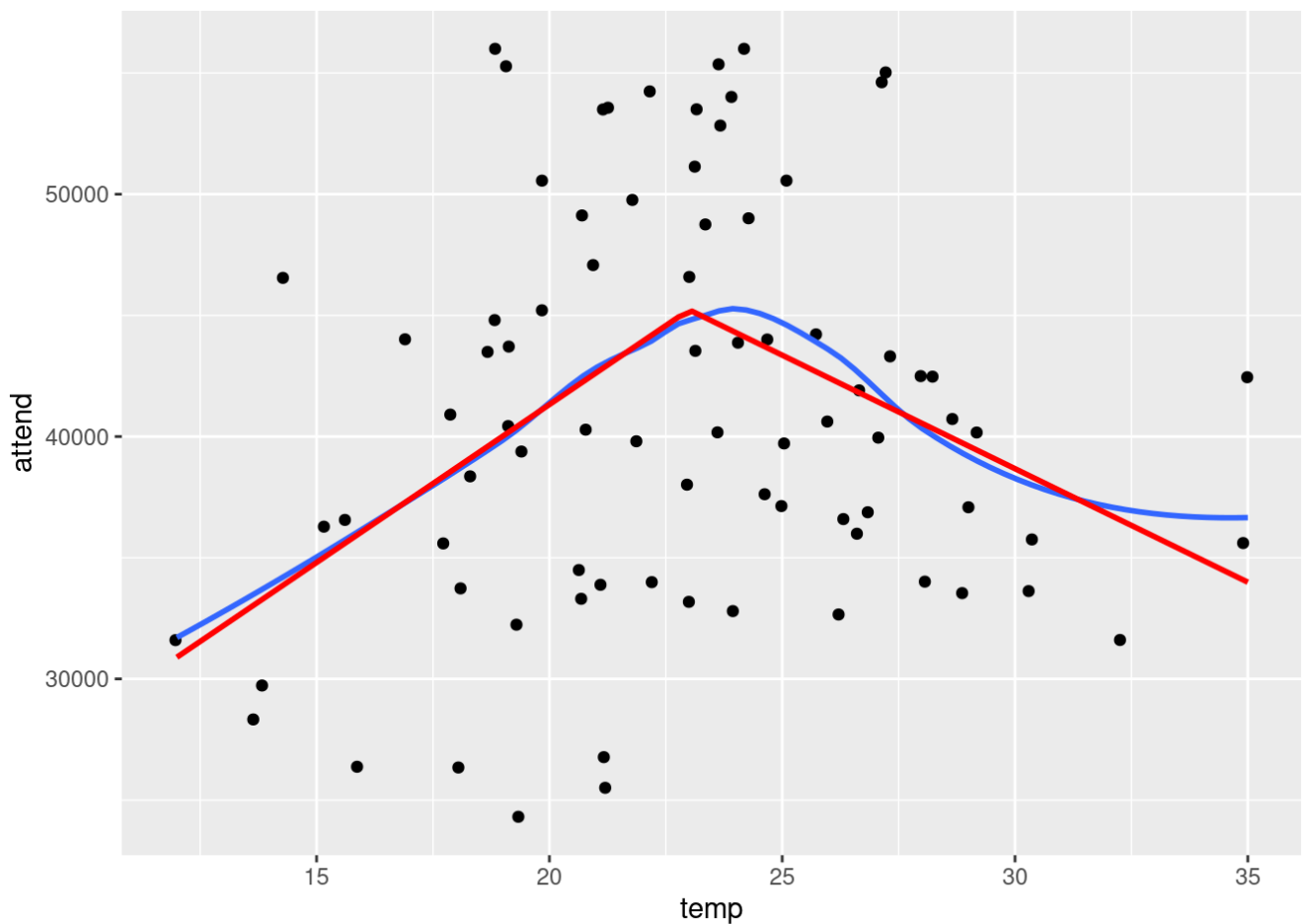
Code

```
##
## Call:
## lm(formula = attend ~ temp + pmax(0, temp - 23), data = events)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17115.9  -5194.3    422.1    4789.0   15982.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15253.3     7363.9   2.071  0.041631 *
## temp           1303.4       360.2   3.618  0.000525 ***
## pmax(0, temp - 23) -2240.1       612.7  -3.656  0.000463 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7727 on 78 degrees of freedom
## Multiple R-squared:  0.1544, Adjusted R-squared:  0.1327
## F-statistic:  7.12 on 2 and 78 DF, p-value: 0.001445
```

$$attend = \beta_0 + \beta_1 temp + \beta_2(temp - 23)^+ + \varepsilon_i$$

Code

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



There is statistically significant relation between attendance and temperature.

## Next time: A linear regression model

Regress attendance on month, day of the week, and bobblehead promotion.

Code

```
## [1] "APR" "MAY" "JUN" "JUL" "AUG" "SEP" "OCT"
```

Code

```
##
## Call:
## lm(formula = attend ~ month + day_of_week + bobblehead, data = event
s2)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-10786.5	-3628.1	-516.1	2230.2	14351.0

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	47796.4	2671.9	17.889	< 0.0000000000000000
monthAPR	-7163.2	2732.7	-2.621	0.01083
monthMAY	-9548.9	2526.9	-3.779	0.00033
monthJUL	-4313.4	2767.9	-1.558	0.12386
monthAUG	-4785.3	2594.6	-1.844	0.06955
monthSEP	-7134.2	2763.5	-2.582	0.01202
monthOCT	-7825.9	4232.6	-1.849	0.06887
day_of_weekMonday	-6724.0	2506.7	-2.682	0.00919
day_of_weekTuesday	1187.5	2594.7	0.458	0.64867
day_of_weekWednesday	-4264.0	2501.4	-1.705	0.09289
day_of_weekThursday	-5948.6	3339.3	-1.781	0.07938
day_of_weekFriday	-1840.2	2426.8	-0.758	0.45094
day_of_weekSaturday	-351.9	2417.6	-0.146	0.88469
bobbleheadYES	10714.9	2419.5	4.429	0.000035

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6120 on 67 degrees of freedom
## Multiple R-squared:  0.5444, Adjusted R-squared:  0.456
## F-statistic: 6.158 on 13 and 67 DF, p-value: 0.0000002083
```

8. Is there any evidence for a relationship between attendance and other variables? Why or why not?

Code

```
##
## Call:
## lm(formula = attend ~ month + day_of_week + bobblehead, data = event
s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10786.5  -3628.1  -516.1   2230.2  14351.0
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|
t|)
## (Intercept)      33909.16    2521.81  13.446 < 0.0000000000000000
2 ***
## monthMAY          -2385.62    2291.22  -1.041      0.3015
2
## monthJUN           7163.23    2732.72   2.621      0.0108
3 *
## monthJUL           2849.83    2578.60   1.105      0.2730
3
## monthAUG           2377.92    2402.91   0.990      0.3259
3
## monthSEP            29.03    2521.25   0.012      0.9908
5
## monthOCT          -662.67    4046.45  -0.164      0.8704
1
## day_of_weekTuesday  7911.49    2702.21   2.928      0.0046
6 **
## day_of_weekWednesday 2460.02    2514.03   0.979      0.3313
4
## day_of_weekThursday  775.36    3486.15   0.222      0.8246
7
## day_of_weekFriday   4883.82    2504.65   1.950      0.0553
7 .
## day_of_weekSaturday 6372.06    2552.08   2.497      0.0150
0 *
## day_of_weekSunday   6724.00    2506.72   2.682      0.0092
0 **
## bobbleheadYES      10714.90    2419.52   4.429      0.000035
9 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6120 on 67 degrees of freedom
## Multiple R-squared:  0.5444, Adjusted R-squared:  0.456
## F-statistic: 6.158 on 13 and 67 DF, p-value: 0.0000002083
```

Check F-statistic's p-value. If it is less than 0.05, then there is relation between attendance and predictors.

9. Does the `bobblehead` promotion have a statistically significant effect on the attendance?

Test  $H_0 : \beta_{\text{BobbleheadYES}} = 0$ . Under  $H_0$ , t-stat in the summary has t-distribution with degrees of freedom equal to (number of samples - numbr of parameters estimated). We check directly p-value for the t-test. If p-value is small ( $< 0.05$ ), then we reject the null hypothesis and conclude that BobbleHead is important in increasing the attendance in the games. Since p-value (0.0000359) is less than 5%, we reject the null. Therefore we conclude it is a good idea to use bobblehead to boost the number of fans coming to stadium to watch the game.

10. Do `month` and `day of week` variables help to explain the number of attendants?

Is there a relation between month and attendance (after we account for the effects of `day_of_week` and `bobblehead`)?

Code

```
## Analysis of Variance Table
##
## Model 1: attend ~ day_of_week + bobblehead
## Model 2: attend ~ month + day_of_week + bobblehead
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1      73 3129721926
## 2      67 2509574563   6 620147363 2.7594 0.01858 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0$ : the small model is correct. If p-value is small (for example, less than 5%), as always we reject the null hypothesis (in this case, null says that the small model is correct). Here, p-value = 0.01858  $<$  5%, so it is small. We reject the small model. Therefore, we conclude that month and attendance are related (while `day_of_week` and `bobbleheadYES` are still in the model).

11. How many fans are expected to be drawn alone by a bobblehead promotion to a home game?

Give a 90% confidence interval.

12. How good does the model fit to the data? Why? Comment on residual standard error and  $R^2$ . Plot observed attendance against predicted attendance.

Is day of week important? (Does `day_of_week` provide new explanation while the other predictors are still present in the model?)

Code

```
## Analysis of Variance Table
##
## Model 1: attend ~ month + bobblehead
## Model 2: attend ~ month + day_of_week + bobblehead
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1      73 3085413762
## 2      67 2509574563   6 575839199 2.5623 0.02704 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the small model because p-value (0.02704) is small (less than 0.05). SO conclude that day of week still contributes important information to pur understanding of attendance while the others two predcitors are in the model.

Variable selection

Code

```
## Analysis of Variance Table
##
## Model 1: attend ~ bobblehead
## Model 2: attend ~ month + bobblehead
## Model 3: attend ~ month + day_of_week + bobblehead
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1      79 3642937151
## 2      73 3085413762   6 557523389 2.4808 0.03157 *
## 3      67 2509574563   6 575839199 2.5623 0.02704 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
##
## Call:
## lm(formula = attend ~ month + bobblehead, data = events)
##
## Coefficients:
##   (Intercept)      monthMAY      monthJUN      monthJUL      mon
thAUG
##      38519.6      -2603.6      6561.4      2147.8      1
658.4
##      monthSEP      monthOCT  bobbleheadYES
##      435.4      -1816.0      12867.3
```

Code



```
## Analysis of Variance Table
##
## Model 1: attend ~ bobblehead
## Model 2: attend ~ day_of_week + bobblehead
## Model 3: attend ~ month + day_of_week + bobblehead
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      79 3642937151
## 2      73 3129721926   6 513215226 2.2836 0.04583 *
## 3      67 2509574563   6 620147363 2.7594 0.01858 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## Analysis of Variance Table
##
## Model 1: attend ~ bobblehead
## Model 2: attend ~ month + day_of_week + bobblehead
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      79 3642937151
## 2      67 2509574563  12 1133362588 2.5215 0.008408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- It is wise to compare all nested models pairwise with anova()
- If p-value is slightly above 5%, we may call anova to be inconclusive. Use cross-validation to decide between two models:
  - For cross-validation:
    - Split the data into folds (say 10 for large data, 5 or small for small data; here I would have taken 5)
    - For each of five folds
      - Remove the test fold
      - Train all models that you want to compare on the remaining four folds combined.
      - Test all models on the held-out fold
        - Calculate MSE, MAE
    - Take average of performance measures (MSE, MAE) across all ftest folds.
    - Pick the model which has the most favorable performance measures.

13. Predict the number of attendees to a typical home game on a Wednesday in June if a bobblehead promotion is extended. Give a 90% prediction interval.

We will use the full model because anova analysis showed that all predictors were important

Code

```
##           fit      lwr      upr
## 1 54247.32 42491.1 66003.55
```

## Project (will be graded)

Include **all variables** and conduct a full regression analysis of the problem. Submit your R markdown and html files to course homepage on moodle.

- The project will be due 19:00 on Saturday, April 9, 2022.
- You can form groups with at most three members. Submit one report for the whole group.
- Submit Rmd and html in a single zip named with student ids of group members to Moodle page. There will soon be a link on the Moodle page. Do not forget to write names of group members, student ids inside Rmd.

## Bibliography

Baumer, B. S., D. T. Kaplan, and N. J. Horton. 2017. *Modern Data Science with R*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. <https://books.google.com.tr/books?id=NrddDgAAQBAJ> (<https://books.google.com.tr/books?id=NrddDgAAQBAJ>).

Miller, T. W. 2014. *Modeling Techniques in Predictive Analytics with Python and R: A Guide to Data Science*. FT Press Analytics. Pearson Education. <https://books.google.com.tr/books?id=PU6nBAAAQBAJ> (<https://books.google.com.tr/books?id=PU6nBAAAQBAJ>).

Wickham, H., and G. Grolemund. 2017. *R for Data Science*. O'Reilly Media. <https://books.google.com.tr/books?id=aZRYrgEACAAJ> (<https://books.google.com.tr/books?id=aZRYrgEACAAJ>).