

Rickord, Jake JJ

2021-09-16

## Exercise 3.2 Using Data to Improve a Marketing Promotion Report

For this specific exercise, our goal was to assist the LA Dodgers (ew) to improve their marketing promotion, specifically targeting a night that would be best to run said promotion to increase attendance. We looked to answer this question via Python by way of Jupyter Notebook.

Delving into our project, we first imported the dodgers.csv dataset and printed the output to examine more closely. Basing our analysis off the specific question we were asked, we recognized that the dataset included day game data. Since the details requested were centered on what night would be best for the promotion, we eliminated the day entries as auxiliary data.

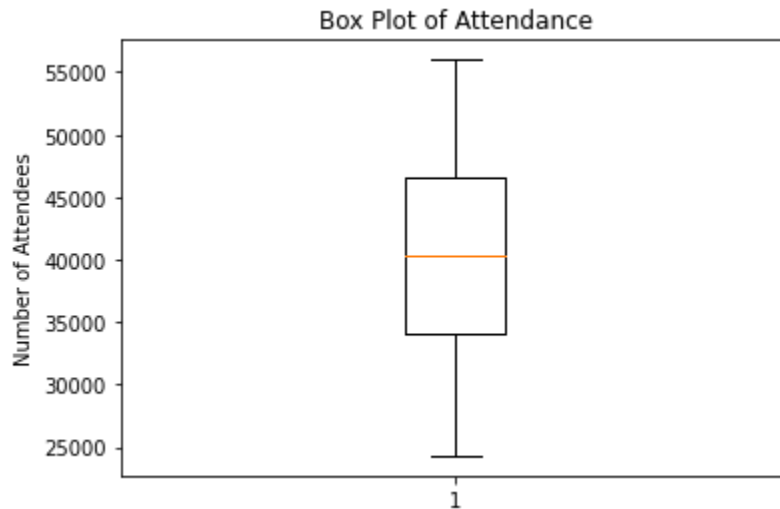
From there, we quantified the marketing promotion data which was subsetting into 4 different types: caps, shirts, fireworks, and a bobblehead. Again, referencing the question being asked, it did not request the type of promotion to be targeted. As such, we combined the columns together, finding any promotion being offered on a night to be considered a TRUE, or 1 case, and nights where no promotions offered identified as FALSE or 0 value.

Once we got this information set up, we were able to start digging into the attendance data. To do so, we first examined measures of central tendency, and plotted the interquartile range of this variable to identify normal attendance behavior:

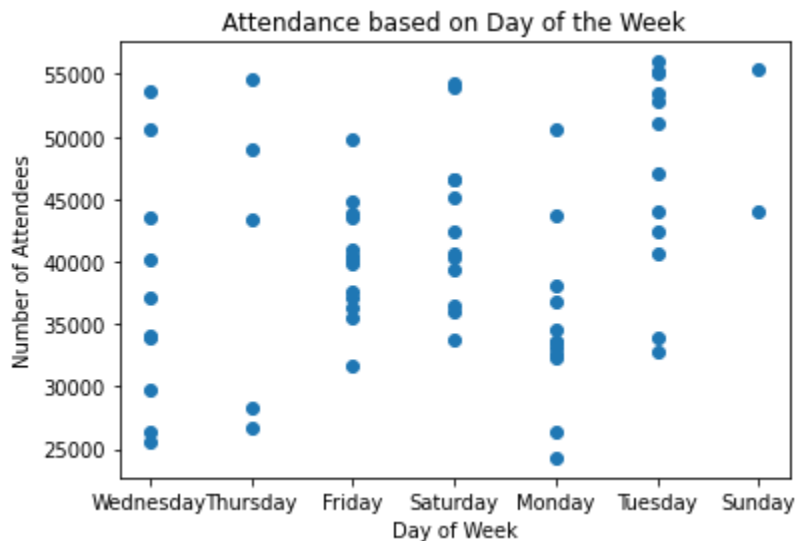
```
count    66.000000
mean    40868.893939
std      8551.130331
min     24312.000000
25%     33998.250000
50%     40358.000000
75%     46578.250000
max      56000.000000
Name: attend, dtype: float64
```

Judging from our plots, most industries have annual salaries floating just under \$50,000, which is actually quite surprising. Very small crease through the \$100,000 marker, with few outliers wait up there, and a small amount near \$0. This is backed up by the following stats:

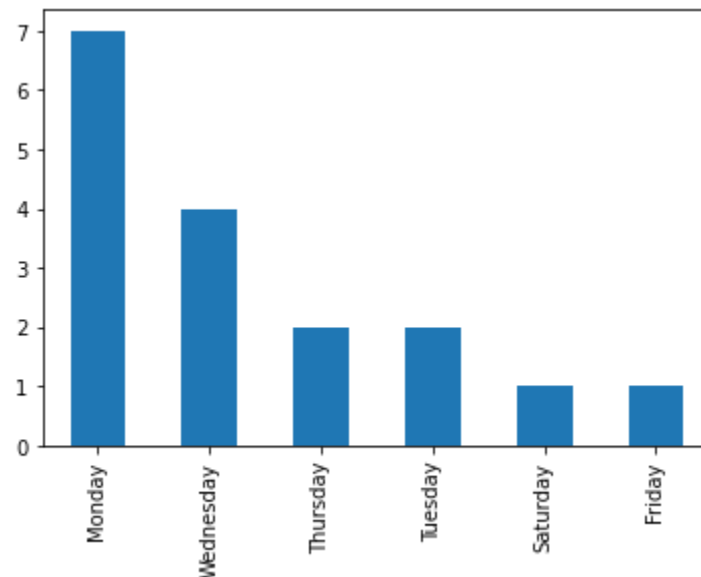
|       |               |
|-------|---------------|
| count | 1329.000000   |
| mean  | 56712.919488  |
| std   | 28260.586899  |
| min   | 0.000000      |
| 25%   | 36810.000000  |
| 50%   | 49520.000000  |
| 75%   | 69600.000000  |
| max   | 207380.000000 |



From this plot and these statistics, we identified that outliers would fall under attendance numbers of ~34,000 persons or above ~46,600 persons. For our first analysis, we're not concerned with those outliers that outperformed the IQR. To identify these outliers, we review the following scatter plot:



Particular points of interest are those outlier points that fall under the IQR, of which all but Sunday seem to have at least one. As such, we'd like to see which day tends to have the lowest attendance on a regular basis. The key to doing so is comparing the count of games per day that had attendance under that IQR, so we subsetting our dataset to those instances, and plotted a bar graph for easy comparison:



Without any further analysis, we'd say Monday would be the optimal day to target since this has, on average, the most days with low attendance, which makes sense given it's the start of the work week. However, we took this one step further, and leveraged regression analysis to identify the days that had the highest impact from marketing promotions to see the most effective deployments of marketing promotions. The results of linear regression models for each day are shown below:

- coefficient of determination for Wednesday is 0.3500624054817145
  - intercept is 35655.666666666664
  - slope is [17914.33333333]
- coefficient of determination for Thursday is 0.7037820225760137
  - intercept is 32803.333333333336
  - slope is [19010.16666667]
- coefficient of determination for Friday is 0.0
  - intercept is 40116.92307692308
  - slope is [0.]
- coefficient of determination for Saturday is 0.6211489610725576
  - intercept is 40747.5
  - slope is [13380.5]
- coefficient of determination for Monday is 0.4912909324769106
  - intercept is 33548.090909090904
  - slope is [17010.90909091]

- coefficient of determination for Tuesday is 0.1188413196432272
  - intercept is 43196.0
  - slope is [5785.5]
- coefficient of determination for Sunday is 1.0
  - intercept is 44005.0
  - slope is [11354.]

Judging by these regression values (with Sunday and Fridays being oddballs due to Sunday only having two samples and Fridays having no records without a marketing promotion), Thursday would be the most efficient use of marketing promotions to increase attendance.

Overall, our goal of answering the objective of “determine what night would be the best to run a marketing promotion to increase attendance”, we would likely request further information to determine what they mean by “best”, however if I had to hazard to guess, they’re looking for the most efficient deployment of marketing promotions, which lends itself to the Thursday model identified in our regression analysis modeling.