Impact of Musicians' Spending and Proximity on Number of Paid Gigs

Benny Nguyen, Ananya Venkateswaran

Elements of Data Science, Spring 2024

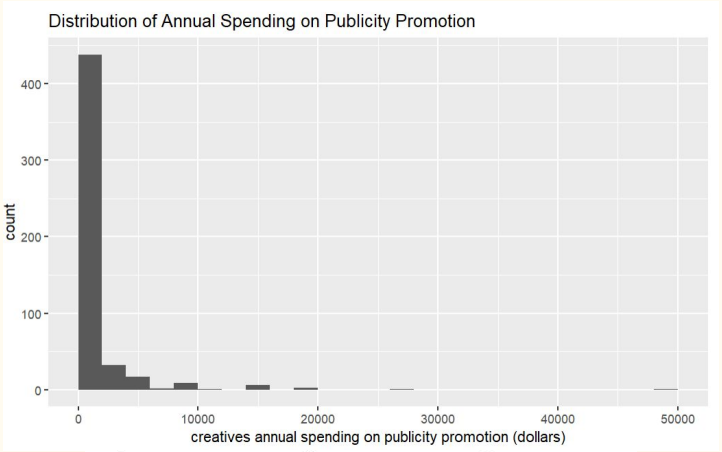
Introduction

- Motivation behind choice
  - ATX music capital, diverse scene
- How does proximity to downtown impact a musicians success (increased number of performances)?
  - Testing to observe if fitted regression model for proximity to downtown and number of paid performances is accurate to the dataset
  - Univariate and bivariate graphs are included that showcase each distribution
- How does spending affect a musician's number of paid performances?
  - Testing to observe if fitted regression model for spending and number of paid performances is accurate to the dataset
  - Univariate and bivariate graphs are included that showcase each distribution

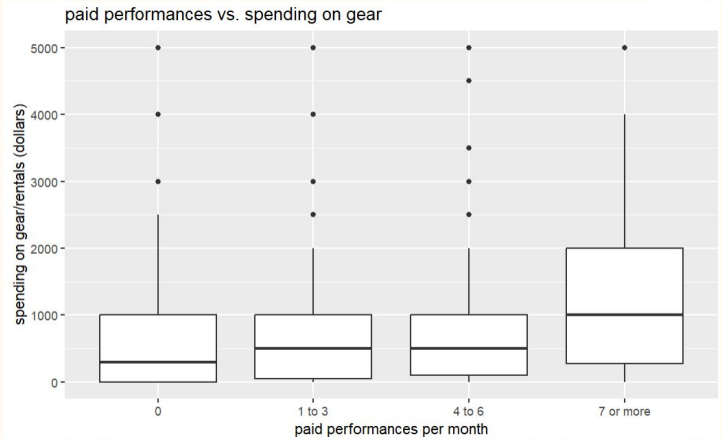
Methods

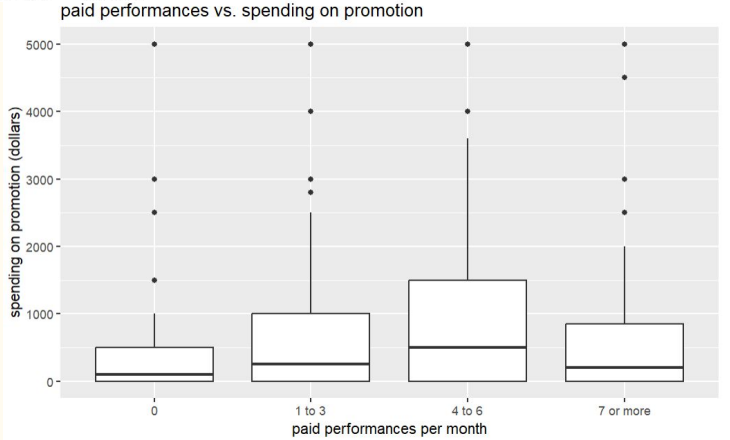
- We started off with 107 columns and 2207 rows.
- We had to narrow this down into 6 columns (variables) and 510 rows.
- First we filtered out our variables of interest. Then, we removed all rows that had missing values for any of these variables. Finally, we had to recode our categorical variable "creatives paid performances per month currently" in order to regroup the variable into 4 categories.
- Next, a k-nearest neighbors model was run using the knn3() function in RStudio
    - Categorical outcome variable
- Efficacy was analyzed by finding the performance of each k-fold and and then observing average performance over all k-folds

EDA 1




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0       0     250    1399    1000   50000
```





median(`CREATIVES Annual Spending on Gear or Rentals`) <dbl>	IQR(`CREATIVES Annual Spending on Gear or Rentals`) <dbl>
300	1125
500	930
500	900
1000	1700

median(`CREATIVES Annual Spending on Publicity Promotion`) <dbl>	IQR(`CREATIVES Annual Spending on Publicity Promotion`) <dbl>
200	625
250	1000
850	1625
200	1000

Model 1

```
fit_knn <- knn3(performances ~ publicity + merch + gear,
                data = music_data,
                k = 5)

predict(fit_knn, music2) |> as.data.frame() |> head()

music_pred <- music_data |>
  mutate(prob_class1 = predict(fit_knn, newdata = music_data)[,1],
         prob_class2 = predict(fit_knn, newdata = music_data)[,2],
         prob_class3 = predict(fit_knn, newdata = music_data)[,3],
         prob_class4 = predict(fit_knn, newdata = music_data)[,4]) |>
  group_by(ID) |>
  mutate(pred_class = case_when(
           prob_class1 >= max(prob_class2, prob_class3, prob_class4) ~ "0",
           prob_class2 >= max(prob_class1, prob_class3, prob_class4) ~ "1 to 3",
           prob_class3 >= max(prob_class1, prob_class2, prob_class4) ~ "4 to 6",
           prob_class4 >= max(prob_class1, prob_class2, prob_class3) ~ "7 or more"
         ))

mean(music_pred$performances == music_pred$pred_class, na.rm = TRUE)
```

	0 <dbl>	1 to 3 <dbl>	4 to 6 <dbl>	7 or more <dbl>
1	0.20000…	0.40000…	0.40000…	0.00000…
2	0.20000…	0.40000…	0.20000…	0.20000…
3	0.00000…	0.50000…	0.16666…	0.33333…
4	0.00000…	0.50000…	0.16666…	0.33333…
5	0.14285…	0.28571…	0.00000…	0.57142…
6	0.00000…	0.75000…	0.12500…	0.12500…

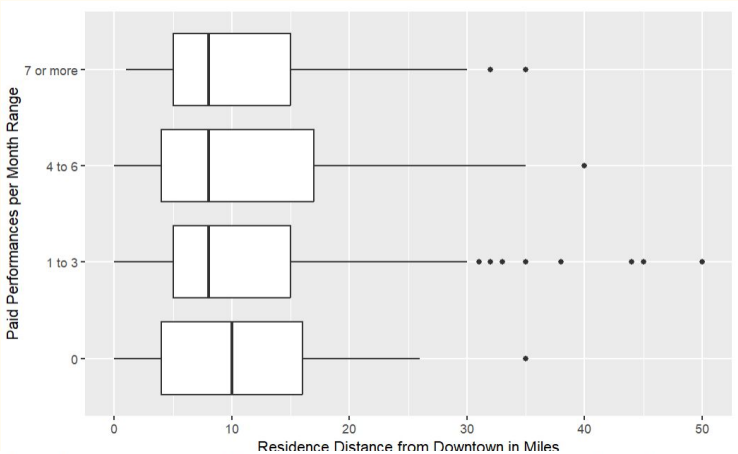
```
[1] 0.6215686
```

CV 1

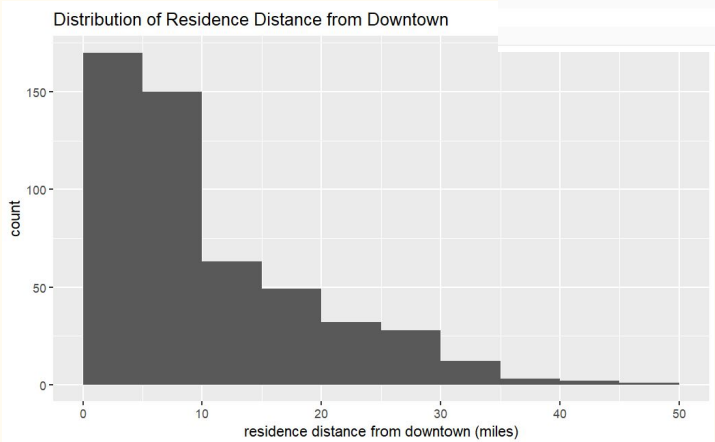
```
perf_k <- NULL

for(i in 1:k){
  train_not_i <- data[folds != i, ]
  test_i <- data[folds == i, ]

  train_model <- knn3(performances ~ publicity + merch + gear,
                      data = train_not_i,
                      k = 5)

  test_i <- test_i |>
  mutate(prob_class1 = predict(train_model, newdata = test_i)[,1],
         prob_class2 = predict(train_model, newdata = test_i)[,2],
         prob_class3 = predict(train_model, newdata = test_i)[,3],
         prob_class4 = predict(train_model, newdata = test_i)[,4]) |>
  group_by(ID) |>
  mutate(pred_class = case_when(
           prob_class1 >= max(prob_class2, prob_class3, prob_class4) ~ "0",
           prob_class2 >= max(prob_class1, prob_class3, prob_class4) ~ "1 to 3",
           prob_class3 >= max(prob_class1, prob_class2, prob_class4) ~ "4 to 6",
           prob_class4 >= max(prob_class1, prob_class2, prob_class3) ~ "7 or more"
         ))
  perf_k[i] <- mean(test_i$performances == test_i$pred_class, na.rm = TRUE)
}

perf_k
mean(perf_k)
sd(perf_k)
```

```
[1] 0.5000000 0.5196078 0.5000000 0.5490196 0.5392157
[1] 0.5215686
[1] 0.02235638
```

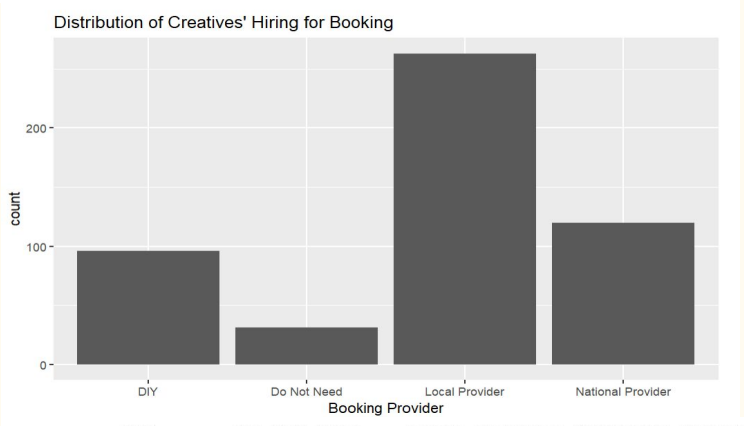
EDA 2



mean(`Residence Distance from Downtown in miles`) <dbl>	sd(`Residence Distance from Downtown in miles`) <dbl>
10.86364	8.177083
11.45734	9.370369
11.29545	9.676561
11.29412	8.281971




```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    5.00    8.00   11.35   15.00   50.00
```




```
       DIY     Do Not Need  Local Provider National Provider
        96              31             263              120
```

Model 2

```
fit_knn3 <- knn3(performances ~ Distance + Hiring, data = music_data, k = 5)

predict(fit_knn3, music_data) |> as.data.frame() |> head()

music_pred <- music_data |>
  mutate(prob_class1 = predict(fit_knn3, newdata = music_data)[,1],
         prob_class2 = predict(fit_knn3, newdata = music_data)[,2],
         prob_class3 = predict(fit_knn3, newdata = music_data)[,3],
         prob_class4 = predict(fit_knn3, newdata = music_data)[,4]) |>
group_by(ID) |>
  mutate(pred_class = case_when(
         prob_class1 >= max(prob_class2, prob_class3, prob_class4) ~ "0",
         prob_class2 >= max(prob_class1, prob_class3, prob_class4) ~ "1 to 3",
         prob_class3 >= max(prob_class1, prob_class2, prob_class4) ~ "4 to 6",
         prob_class4 >= max(prob_class1, prob_class2, prob_class3) ~ "7 or more"))

mean(music_pred$performances == music_pred$pred_class, na.rm = TRUE)
```

	0 <dbl>	1 to 3 <dbl>	4 to 6 <dbl>	7 or more <dbl>
1	0.800000...	0.20000...	0.00000...	0.00000...
2	0.142857...	0.57142...	0.14285...	0.14285...
3	0.000000...	0.37500...	0.37500...	0.25000...
4	0.000000...	0.33333...	0.50000...	0.16666...
5	0.000000...	0.80000...	0.00000...	0.20000...
6	0.047619...	0.61904...	0.09523...	0.23809...

```
[1] 0.5960784
```

CV 2

```
perf_k <- NULL

for(i in 1:k){
  train_not_i <- data[folds != i, ]
  test_i <- data[folds == i, ]

  train_model <- knn3(performances ~ Distance + Hiring,
                      data = train_not_i,
                      k = 5)

  test_i <- test_i |>
  mutate(prob_class1 = predict(train_model, newdata = test_i)[,1],
         prob_class2 = predict(train_model, newdata = test_i)[,2],
         prob_class3 = predict(train_model, newdata = test_i)[,3],
         prob_class4 = predict(train_model, newdata = test_i)[,4]) |>
  group_by(ID) |>
mutate(pred_class = case_when(
         prob_class1 >= max(prob_class2, prob_class3, prob_class4) ~ "0",
         prob_class2 >= max(prob_class1, prob_class3, prob_class4) ~ "1 to 3",
         prob_class3 >= max(prob_class1, prob_class2, prob_class4) ~ "4 to 6",
         prob_class4 >= max(prob_class1, prob_class2, prob_class3) ~ "7 or more")
         )
  perf_k[i] <- mean(test_i$performances == test_i$pred_class, na.rm = TRUE)
}

perf_k
mean(perf_k)
sd(perf_k)
```

```
 [1] 0.5196078 0.5686275 0.5686275 0.5882353 0.5098039
 [1] 0.5509804
 [1] 0.03424363
```

Results

- In general a lot of our univariate data is left skewed
- Most musicians prefer local booking services
- Based on visualizations, it seems that there's a positive correlation between spending and performances

- For both models, the average performance was worse than the dataset's overall performance - which could suggest overfitting
- Models don't do as well on new data
- The proximity model performs slightly better than the spending model, but both are pretty poor overall

Discussion

- Main Takeaways: the variables of distance from downtown Austin and spending are not main predictors of number of paid performances
  - This is further backed up with the average for each of the cross-validations being 0.55 for distance from downtown Austin and Hiring for booking as predictors
  - Also backed up with average for each of the cross-validations being 0.52 for publicity, merch, and gear as predictors
- These findings surprised us - given the idea of distance from downtown (hotspot for music) and other services increasing musician visibility
- These findings could motivate individuals from any background and location to pursue music if it is their dream because it is possible to succeed and book paid performances regardless of these other factors
- Ethical issues could be including personal information such as pay and living situation publicly when the individual may not necessarily have consented to wanting it out there even independent of their identity

Contributions + References

- Benny: worked on models and visualizations to answer: “How does spending affect a musician’s number of paid performances?”
- Ananya: worked on models and visualizations to answer: How does proximity to downtown impact a musicians success (increased number of performances)?
- https://www.austinmusiccensus.org/
- https://www.kut.org/austin/2022-02-15/same-as-it-ever-was-musician-pay-for-live-shows-in-austin-hasnt-changed-in-40-years
- It was challenging to work through the code together just because it isn't possible to share files via RStudio. However, in troubleshooting and forming our dataset, we were able to gain a better understanding of the code we learned in class. Furthermore, we learned how to apply our knowledge of code to a new setting.