Benny Nguyen, Ananya Venkateswaran

1) Title and Introduction

Title: Impact of Musicians’ Spending and Proximity on Number of Paid Gigs

Our dataset is from the 2022 Greater Austin Music Census, conducted by Sound Music Cities. The data is the results from the survey given to people in the music industry in the Greater Austin area.

The dataset that is being investigated documents the amount musicians have to pay for publicity, marketing, instruments, and performing in the Austin area. This was chosen as a topic of interest due to the increased popularity of music in the Austin area with Austin city limits, SXSW, and more. According to KUT (cited at end of document), musicians have had to pay around the same amount for live shows for 40 years. This makes it challenging for musicians to make a significant earning to be able to sustain themselves. We were intrigued to learn more about this by analyzing the various factors that involve spending for musicians.

We are interested in predicting the number of paid performances per month based on the following variables: residence distance from downtown, hiring for booking, annual spending on publicity promotion, annual spending on gear/rentals, and annual spending on merchandise. One row represents a survey response.

We expect that positive correlation between spending and performances. We expect musicians who live closer downtown to have more performances. We also expect local booking services to result in the highest number of performances.

Research Question 1: How does spending affect a musician’s number of performances?

Research Question 2: How does proximity to downtown impact a musician’s success (increased number of performances)?

2) Methods

```
music_census_data <- read_csv("2022_MusicCensus_Details_20240306.csv")

head(music_census_data)
```

```
## # A tibble: 6 × 107
##   `County of Residence` Residence Distance from Downtow…¹ City of Austin Resid…²
##   <chr>                                             <dbl> <lgl>
## 1 Bastrop                                              40 FALSE
## 2 Bastrop                                              22 FALSE
## 3 Bastrop                                              35 FALSE
## 4 Bastrop                                              30 FALSE
## 5 Bastrop                                              30 FALSE
## 6 Bastrop                                              30 FALSE
## # ℹ abbreviated names: ¹`Residence Distance from Downtown in miles`,
## #   ²`City of Austin Resident`
## # ℹ 104 more variables: `Primary Music Ecosystem Sector` <chr>,
## #   `Music Business Structure` <chr>, `Work Location` <chr>,
## #   `Years Experience` <chr>, `Music Education` <chr>,
## #   `Community or Business Participation` <chr>,
## #   `Restored Pre Pandemic Workload` <lgl>, …
```

```
print(nrow(music_census_data))
```

```
## [1] 2227
```

```
print(ncol(music_census_data))
```

```
## [1] 107
```

```
music_census_data <- music_census_data |>

  select(`CREATIVES Paid Performances per Month Currently`, `Residence Distance from Downtown in miles`, `CREATIVES Hiring for Booking`, `CREATIVES Annual Spending on Publicity Promotion`, `CREATIVES Annual Spending on Gear or Rentals`, `CREATIVES Annual Spending on Merchandise`) |>

  drop_na() |>

  mutate(`CREATIVES Paid Performances per Month Currently` = recode(`CREATIVES Paid Performances per Month Currently`,
      '7 to 10' = '7 or more',
      '11 to 15' = '7 or more',
      '16 or more' = '7 or more'))

head(music_census_data)
```

```
## # A tibble: 6 × 6
##   CREATIVES Paid Performances pe…¹ Residence Distance f…² CREATIVES Hiring for…³
##   <chr>                                             <dbl> <chr>
## 1 0                                                    22 Local Provider
## 2 0                                                    35 National Provider
## 3 7 or more                                            30 National Provider
## 4 7 or more                                            30 DIY
## 5 7 or more                                            32 National Provider
## 6 1 to 3                                               20 Local Provider
## # ℹ abbreviated names: ¹`CREATIVES Paid Performances per Month Currently`,
## #   ²`Residence Distance from Downtown in miles`,
## #   ³`CREATIVES Hiring for Booking`
## # ℹ 3 more variables: `CREATIVES Annual Spending on Publicity Promotion` <dbl>,
## #   `CREATIVES Annual Spending on Gear or Rentals` <dbl>,
## #   `CREATIVES Annual Spending on Merchandise` <dbl>
```

```
print(nrow(music_census_data))
```

```
## [1] 510
```

```
print(ncol(music_census_data))
```

```
## [1] 6
```

We started off with 107 columns and 2207 rows.

We had to narrow this down into 6 columns (variables) and 510 rows.

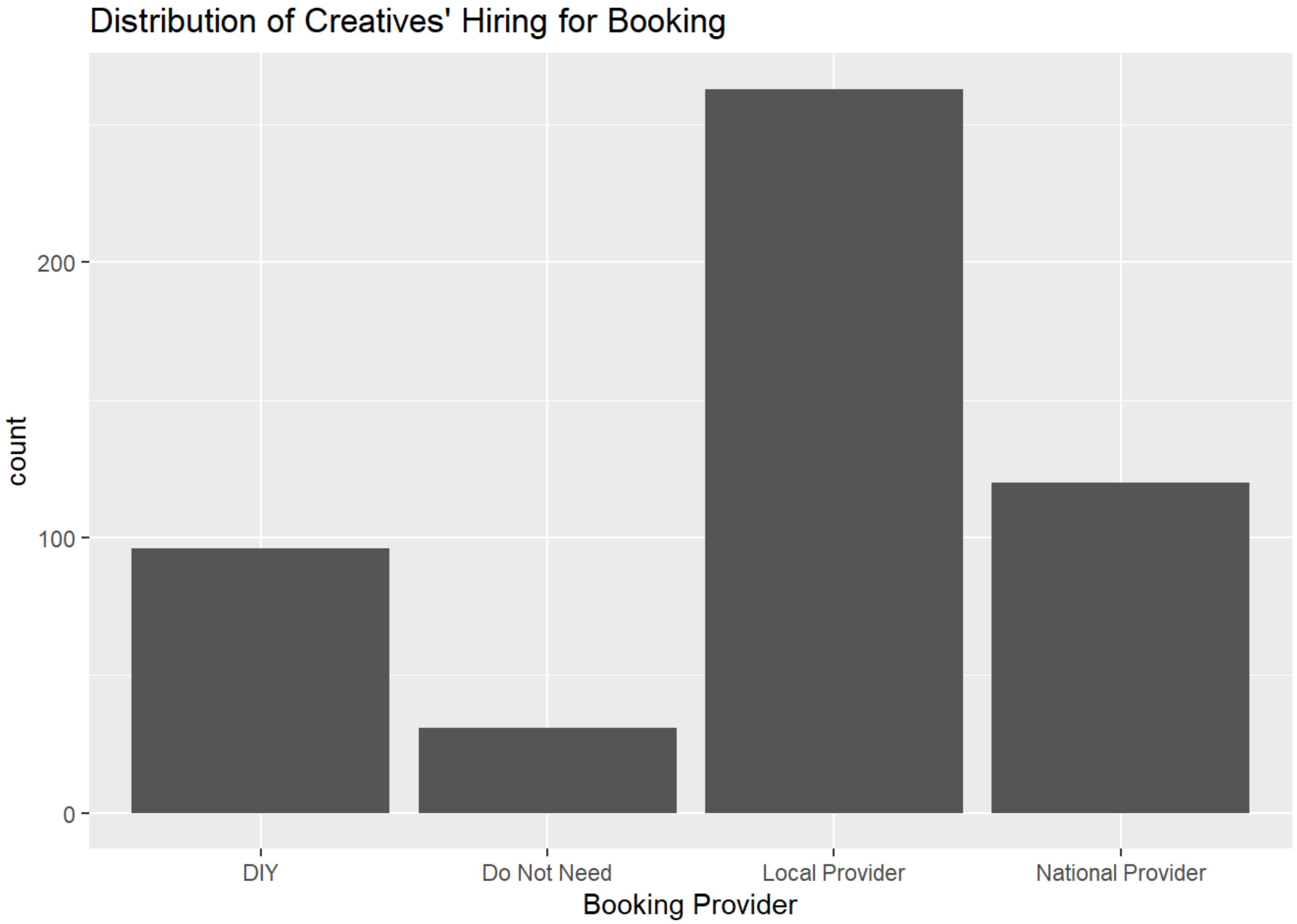
First we filtered out our variables of interest. Then, we removed all rows that had missing values for any of these variables. Finally, we had to recode our categorical variable “creatives paid performances per month currently” in order to regroup the variable into 4 categories.

We believe our resulting dataset is tidy because each variable is defined as a clear column, each observation is a clear row, and they come together to form a table.

3) Results

Visualization 1)

```
music_census_data |>
  ggplot() +
    geom_bar(aes(x = `CREATIVES Hiring for Booking`)) +
      labs(x = "Booking Provider", y = "count", title = "Distribution of Creatives' Hiring
for Booking")
```



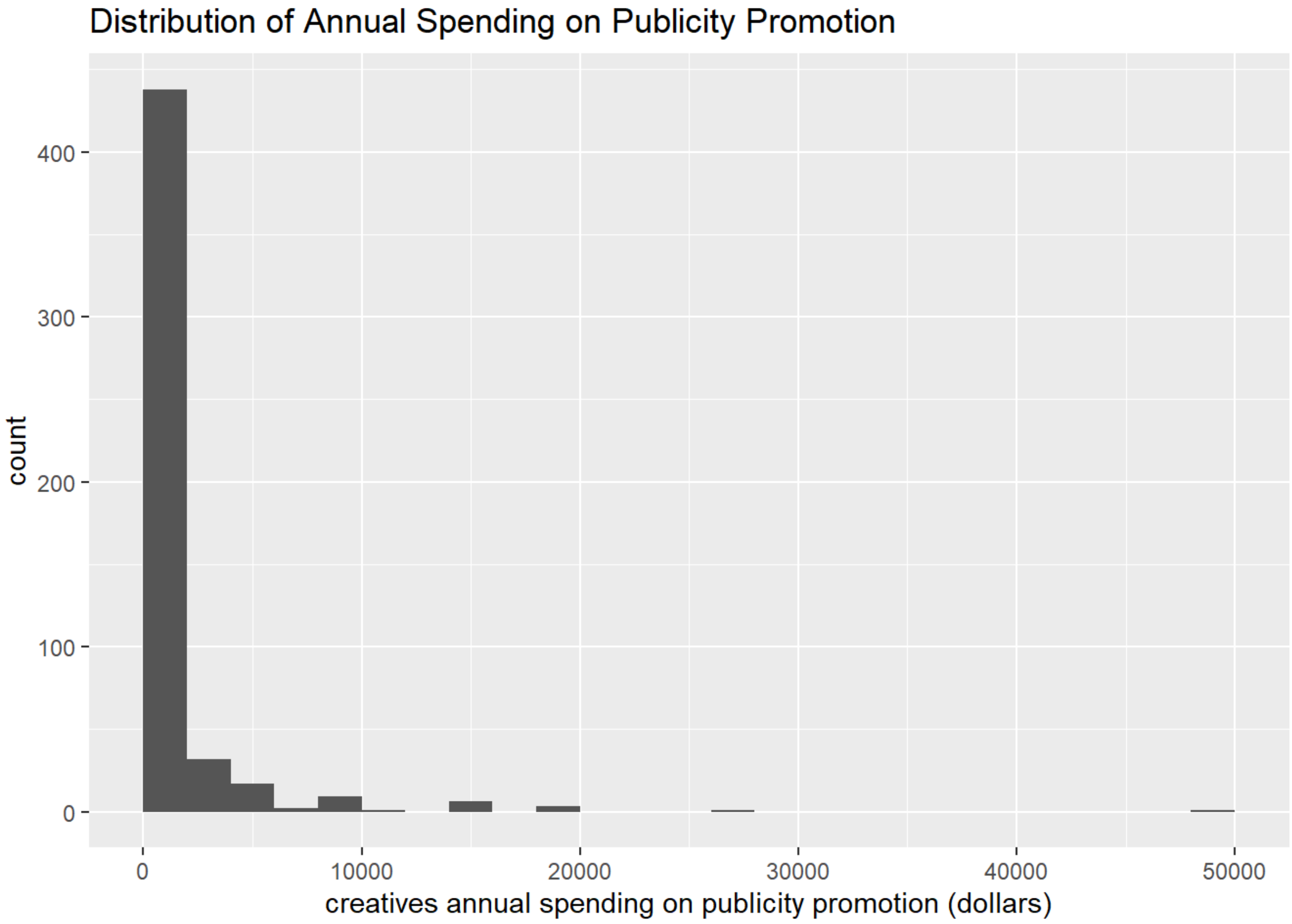
```
table(music_census_data$`CREATIVES Hiring for Booking`)
```

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

Based on our barplot, the most common booking provider is local (mode of 263 responses), followed by national providers , and DIY.

Visualization 2)

```
music_census_data |>
  ggplot() +
    geom_histogram(aes(x = `CREATIVES Annual Spending on Publicity Promotion`), binwidth = 
2000, center = 1000) +
      scale_x_continuous(limits = c(0,50000), breaks = seq(0, 50000, 10000)) +
        labs(x = "creatives annual spending on publicity promotion (dollars)", y = "count", 
title = "Distribution of Annual Spending on Publicity Promotion")
```



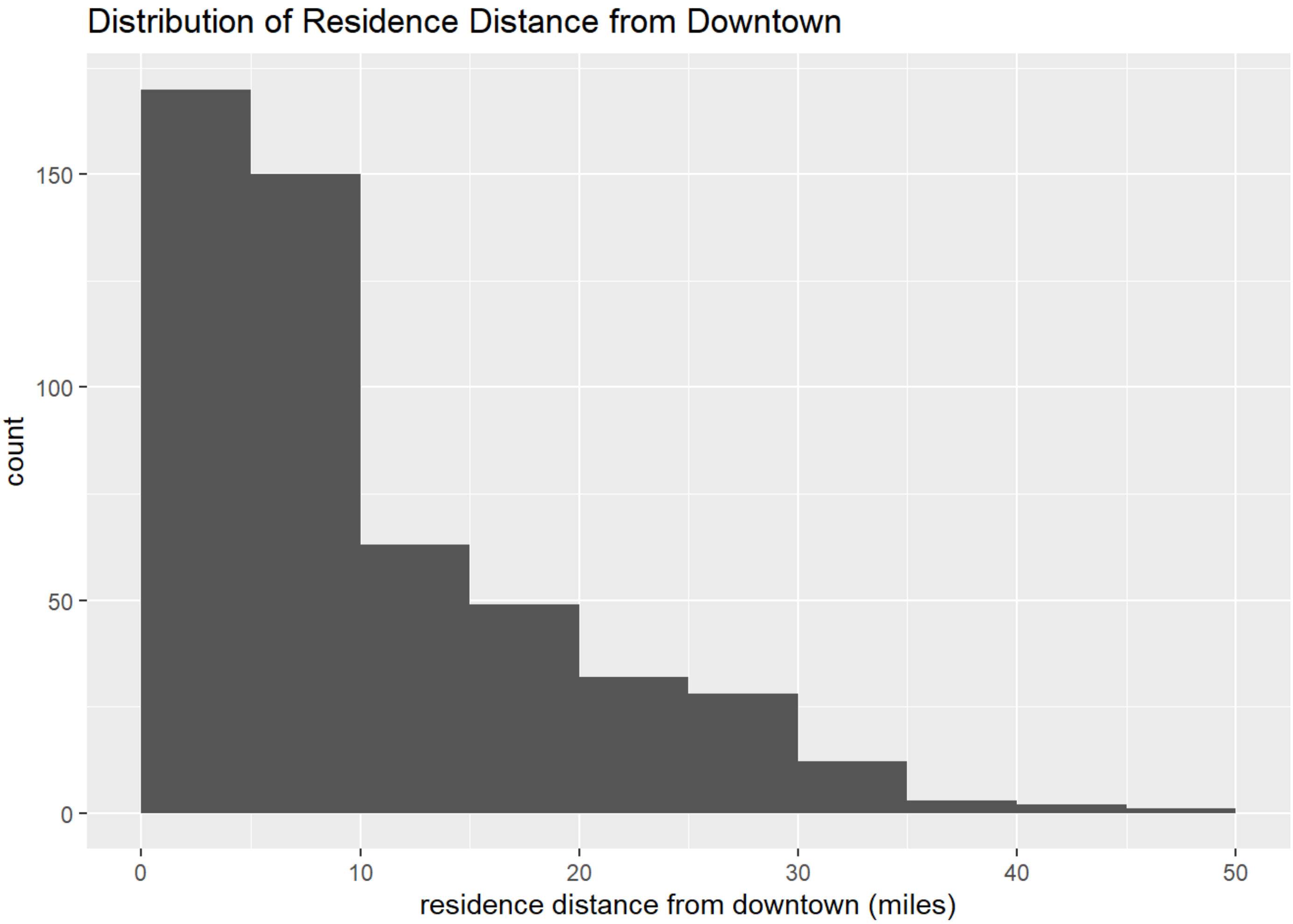
```
summary(music_census_data$`CREATIVES Annual Spending on Publicity Promotion`)
```

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

Based on our histogram, the distribution of spending on publicity promotion seems right skewed. The median is 250 and the IQR is 1000. There do seem to be a couple of outliers that have spendings to up $50000.

Visualization 3)

```
music_census_data |>
  ggplot() +
    geom_histogram(aes(x = `Residence Distance from Downtown in miles`), binwidth = 5, cent
er = 2.5) +
        labs(x = "residence distance from downtown (miles)", y = "count", title = "Distribu
tion of Residence Distance from Downtown")
```



```
summary(music_census_data$`Residence Distance from Downtown in miles`)
```

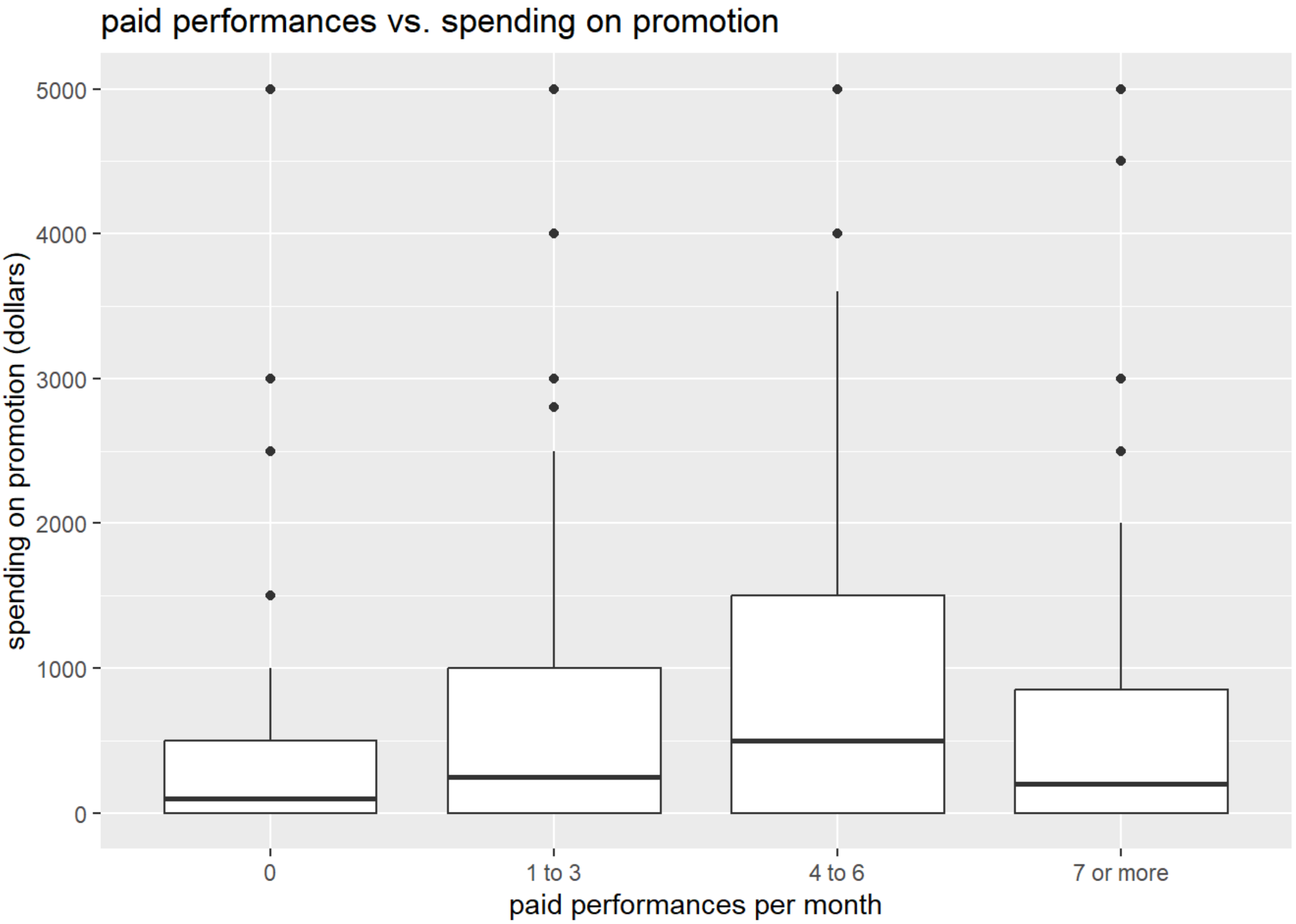
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    5.00    8.00   11.35   15.00   50.00 
```

Based on the histogram, the distribution of distance from downtown is right skewed. The median is 8 miles and the IQR is 10 miles. The range is from 0 to 50 miles.

Visualization 4)

```
music_census_data |>
  ggplot() +
  geom_boxplot(aes(x = `CREATIVES Paid Performances per Month Currently`, y = `CREATIVES An
nual Spending on Publicity Promotion`)) +
  scale_y_continuous(limits = c(0,5000)) +
  labs(y = "spending on promotion (dollars)", x = "paid performances per month", title = "p
aid performances vs. spending on promotion")
```

```
## Warning: Removed 23 rows containing non-finite values (`stat_boxplot()`).
```



```
music_census_data |>
  group_by(`CREATIVES Paid Performances per Month Currently`) |>
  summarize(median(`CREATIVES Annual Spending on Publicity Promotion`),              IQR
(`CREATIVES Annual Spending on Publicity Promotion`))
```

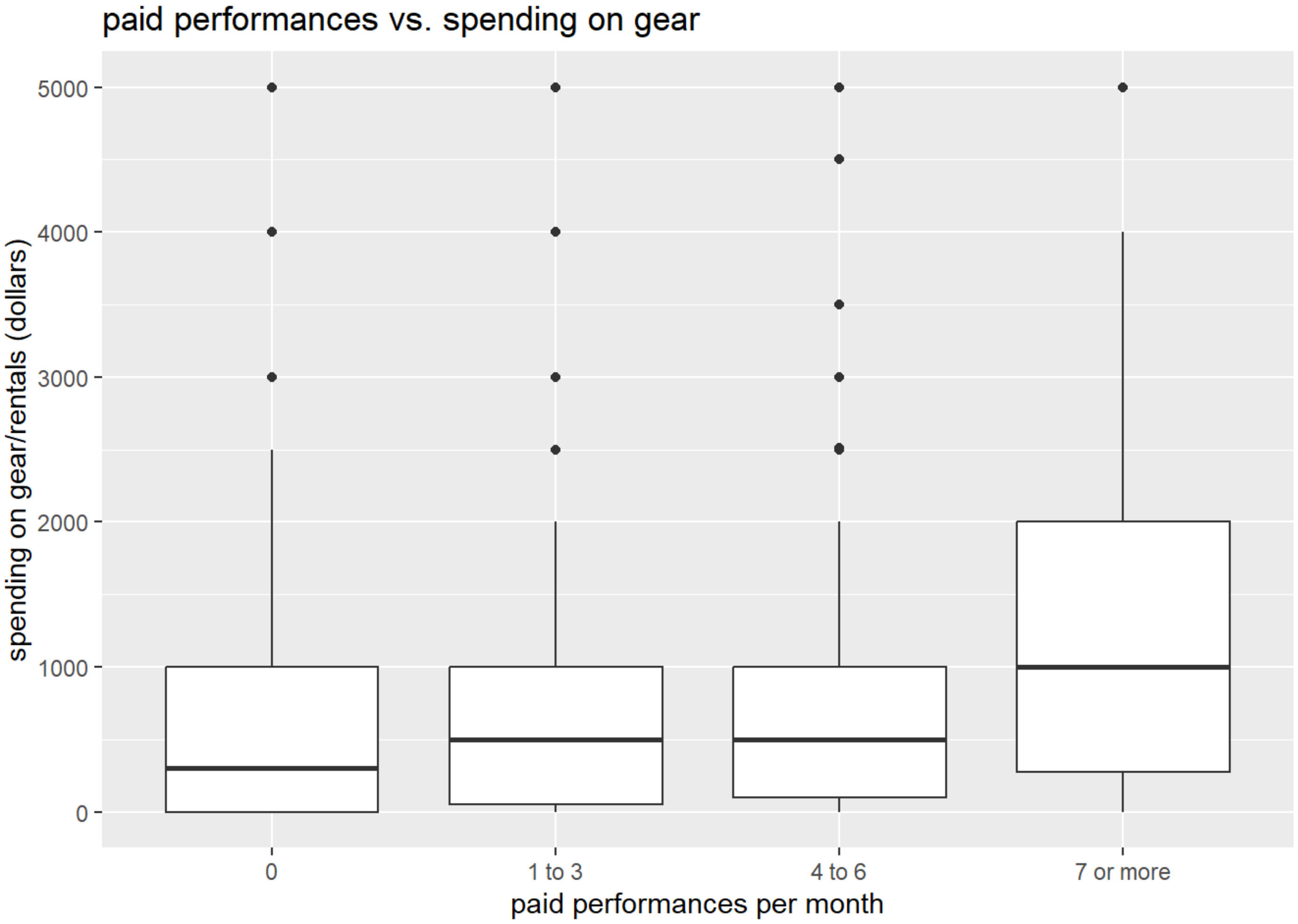
```
## # A tibble: 4 × 3
##   CREATIVES Paid Performances pe…¹ median(\CREATIVES An…² IQR(\CREATIVES Annua…³
##   <chr>                                              <dbl>                  <dbl>
## 1 0                                                    200                    625
## 2 1 to 3                                               250                   1000
## 3 4 to 6                                               850                   1625
## 4 7 or more                                            200                   1000
## # ℹ abbreviated names: ¹`CREATIVES Paid Performances per Month Currently`,
## #   ²`median(\`CREATIVES Annual Spending on Publicity Promotion\`)`,
## #   ³`IQR(\`CREATIVES Annual Spending on Publicity Promotion\`)`
```

Based on the grouped boxplot, there seems to be a positive relationship between spending on promotion and paid performances. As the number of paid performances increase, so does the median spending with the exception of 7 or more performances. 0 performances have a median spending of $200, 1-3 performances have a median of $250, 4-6 performances have a median spending of $850, and 7 or more performances have a median spending of $200. The IQR of 4-6 performances is the greatest ($1625) while the other categories have smaller IQRs. This gives us an answer to spending possibly playing a role in increased performances (positive correlation between the two).

Visualization 5)

```
music_census_data |>
  ggplot() +
    geom_boxplot(aes(x = `CREATIVES Paid Performances per Month Currently`, y = `CREATIVES
Annual Spending on Gear or Rentals`)) +
  scale_y_continuous(limits = c(0,5000)) +
        labs(y = "spending on gear/rentals (dollars)", x = "paid performances per month", t
itle = "paid performances vs. spending on gear")
```

```
## Warning: Removed 14 rows containing non-finite values (`stat_boxplot()`).
```



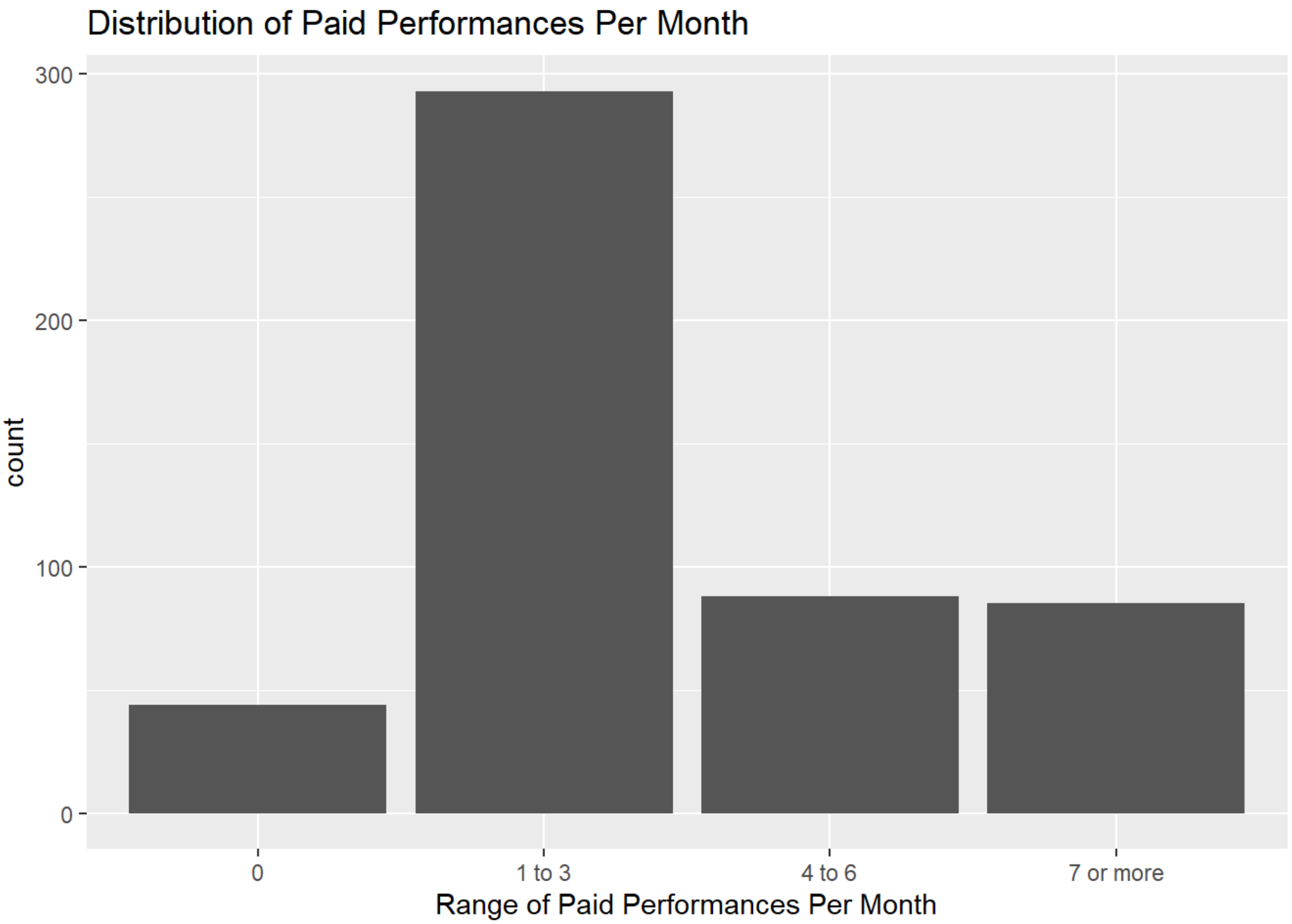
```
music_census_data |>
  group_by(`CREATIVES Paid Performances per Month Currently`) |>
    summarize(median(`CREATIVES Annual Spending on Gear or Rentals`),
              IQR(`CREATIVES Annual Spending on Gear or Rentals`))
```

```
## # A tibble: 4 × 3
##   CREATIVES Paid Performances pe…¹ median(\CREATIVES An…² IQR(\CREATIVES Annua…³
##   <chr>                                             <dbl>                  <dbl>
## 1 0                                                   300                   1125
## 2 1 to 3                                              500                    930
## 3 4 to 6                                              500                    900
## 4 7 or more                                          1000                   1700
## # ℹ abbreviated names: ¹`CREATIVES Paid Performances per Month Currently`,
## #   ²`median(\`CREATIVES Annual Spending on Gear or Rentals\`)`,
## #   ³`IQR(\`CREATIVES Annual Spending on Gear or Rentals\`)`
```

Based on the grouped boxplot, there seems to be a positive relationship between spending on gear and paid performances. As the number of paid performances increase, so does the median spending. 0 performances have a median spending of $300, 1-3 and 4-6 performances have a median spending of $500, and 7 or more performances have a median spending of $1000. The IQR of 7 or more performances is the greatest ($1700) while the other categories have similar IQRs. This gives us an answer to spending possibly playing a role in increased performances (positive correlation between the two).

Visualization 6)

```
# Define the ggplot and the dataframe
ggplot(data = music_census_data) +
  # Use geom_histogram and define mapping aesthetics (we need more!)
  geom_bar(aes(x = `CREATIVES Paid Performances per Month Currently`)) +
  labs(x = "Range of Paid Performances Per Month", y = "count", title = "Distribution of Pa
id Performances Per Month")
```



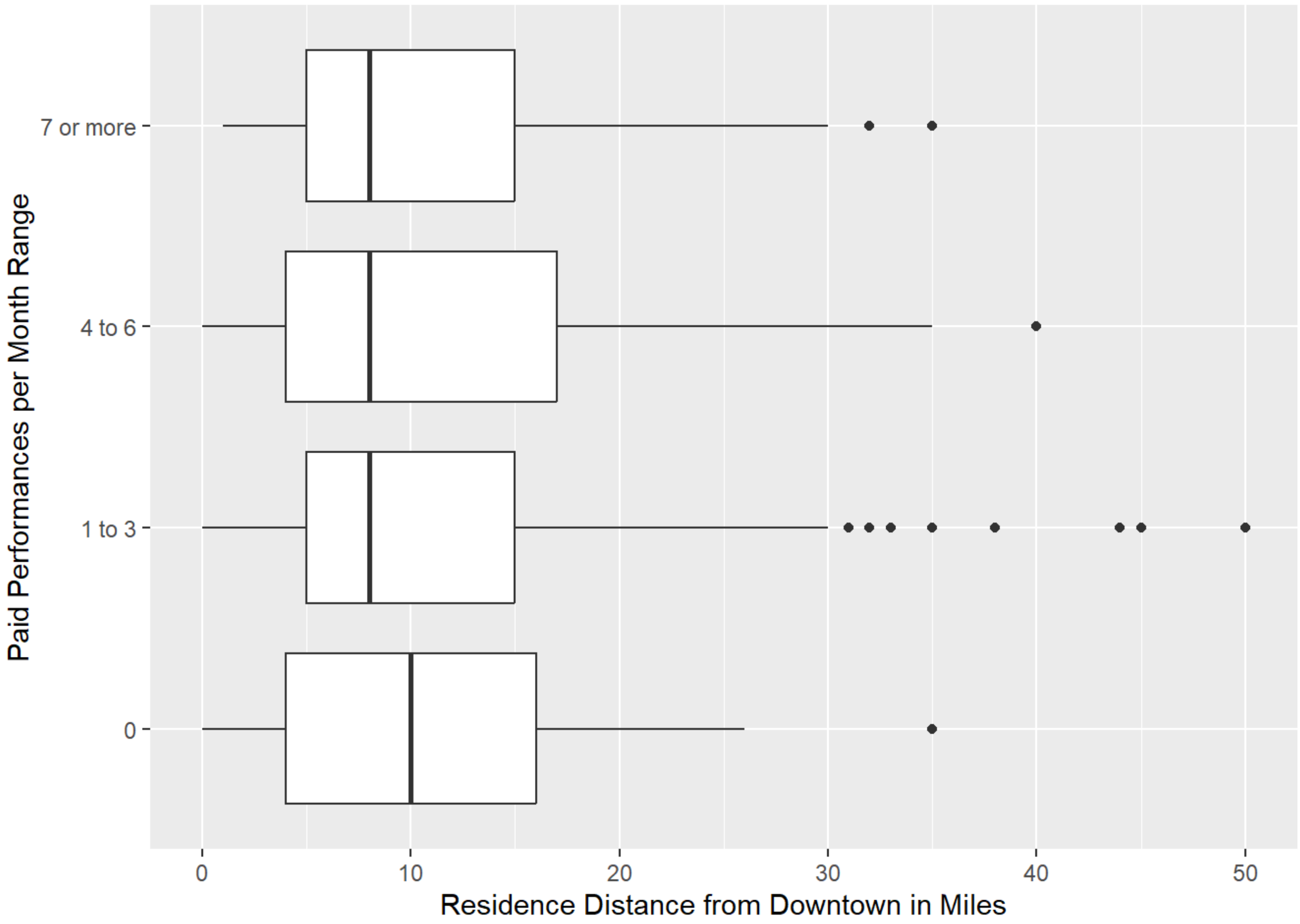
```
#summary statistics
prop.table(table(music_census_data$`CREATIVES Paid Performances per Month Currently`))
```

```
##
##          0     1 to 3     4 to 6  7 or more
## 0.08627451 0.57450980 0.17254902 0.16666667
```

Based on the distribution of paid performances, it is evident most performers are paid for 1-3 performances, with 4-6 and 7+ performances both being the second most common range. This is backed up by the summary statistics, with 57.4% of performers being paid for 1-3 performances and 17.2% and 16.67% being paid for 4-6 performances and 7+ performances, respectively.

Visualization 7)

```
#residence distance from downtown vs. paid performances per month currently
music_census_data |>
  ggplot() +
    geom_boxplot(aes(x = `Residence Distance from Downtown in miles`, y = `CREATIVES Paid Performances per Month Currently`)) +
        labs(y = "Paid Performances per Month Range", x = "Residence Distance from Downtown in Miles")
```



```
music_census_data |>
  # Split the data in groups
  group_by(`CREATIVES Paid Performances per Month Currently`) |>
  # Summarize per group
  summarize(mean(`Residence Distance from Downtown in miles`),
            sd(`Residence Distance from Downtown in miles`))
```

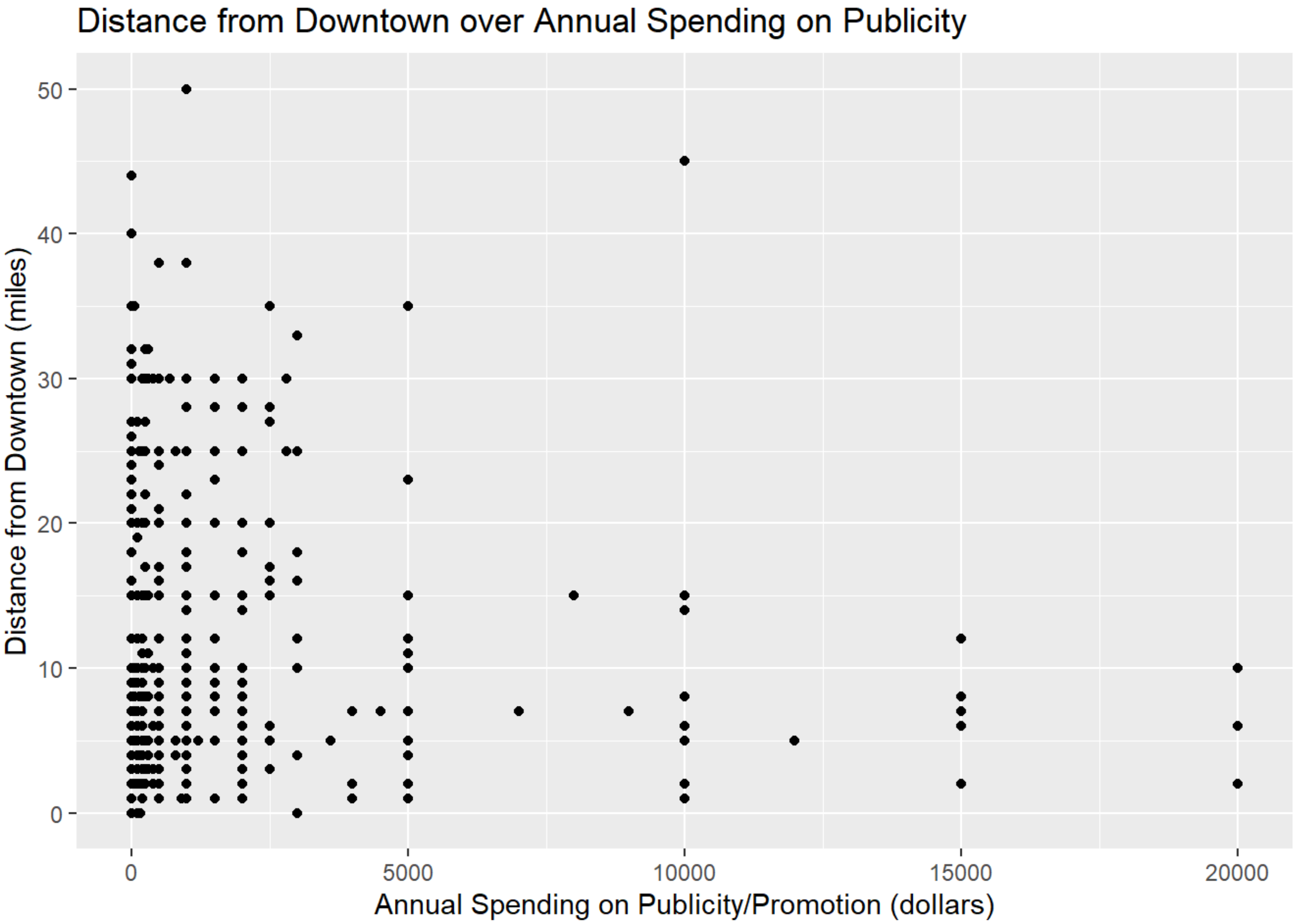
```
## # A tibble: 4 × 3
##   CREATIVES Paid Performances pe…¹ mean(\Residence Dist…² sd(\Residence Distan…³
##   <chr>                                             <dbl>                  <dbl>
## 1 0                                                  10.9                   8.18
## 2 1 to 3                                             11.5                   9.37
## 3 4 to 6                                             11.3                   9.68
## 4 7 or more                                          11.3                   8.28
## # ℹ abbreviated names: ¹`CREATIVES Paid Performances per Month Currently`,
## #   ²`mean(\`Residence Distance from Downtown in miles\`)`,
## #   ³`sd(\`Residence Distance from Downtown in miles\`)`
```

Based on the grouped boxplots, there doesn’t seem to be a significant difference in mean for the ranges 1-3, 4-6, and 7+ paid performances. This is backed up by the summary statistics being around 11 miles for each of the following (same median of 11.26 miles for 4-6 and 7+). The mean for 0 paid performances is a little lower at 10.83 miles. There are many outliers (high distances from downtown) present in each of the ranges, namely in the 1-3 performances range (around 8 far away from downtown but still being paid for 1-3). Each of the ranges except for the “0 group” are right skewed with the 0 group being slightly left skewed. Ths standard deviation for each of the groups are around the same, ranging from 8-9 miles. This answers our second research question: How does proximity to downtown impact a musicians success (increased number of performances)? This data suggests there is not much of a trend/correlation between number of paid performances and distance from downtown Austin.

Visualization 8)

```
music_census_data |>
  ggplot() +
    geom_point(aes(x = `CREATIVES Annual Spending on Publicity Promotion`, y = `Residence D
istance from Downtown in miles`)) +
  scale_x_continuous(limits = c(0,20000)) +
  scale_y_continuous(limits = c(0,50)) +
        labs(x = "Annual Spending on Publicity/Promotion (dollars)", y = "Distance from Dow
ntown (miles)", title = "Distance from Downtown over Annual Spending on Publicity")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



```
cor(music_census_data$`CREATIVES Annual Spending on Publicity Promotion`, music_census_data
$`Residence Distance from Downtown in miles`)
```

```
## [1] -0.05560361
```

There doesn’t seem to be a clear trend between CREATIVES Annual Spending on Publicity and Distance from Downtown Austin. Most individuals seem to be between spending $0-$2500 with a scatter between different distances from downtown. Most distances in this range of spending seem to be between 0-10 miles. The correlation backs up this visual, with a slight negative weak correlation of -0.055. This graph aids us in understanding more about the relationship between distance from downtown Austin and spending. This is important to set the stage for the second research question: How does proximity to downtown impact a musicians success (increased number of performances)?

4) Discussion

On general a lot of our univariate data is left skewed. For spending, most musicians fall under the same general bracket, but there is a fair number of outliers that spend significantly more. In regards to proximity, most musicians live closer to downtown and prefer local booking services.

For our first research question, there seems to be a positive correlation between spending and number of gigs based on visualization 4 and 5. This meets our expectations since investing money should lead to higher quality performances and a broader community outreach. One thing that was surprising is that musicians that get 7 or more monthly gigs actually spend less in promotion than those who get 4-6 gigs.

The visualizations represented above answer our second research question: How does proximity to downtown impact a musicians success (increased number of performances)? We concluded that proximity to downtown does not greatly impact/affect a musician's success. This conclusion was drawn using visualizations 7 and 8. Visualization 7 is a grouped boxplot that depicts distance from downtown Austin (miles) by range of paid performances. The medians depicted in the boxplot are around the same, namely for 1-4, 4-6, 7+ paid performances. The means are also around the same for these (around 11 miles from downtown). The standard deviations for each of the categories are around the same (ranging from 8-9 miles). There also seem to be quite a few outliers in each of the ranges, where there are many individuals who are at least 30 miles from downtown but still book different ranges of performances (specifics described under the visualization). Overall, this randomness in range of paid performance and similarities between groups is evidence to suggest that distance from downtown does not play a significant role in number of paid performances. This was surprising, as we had theorized that closer distance to downtown suggests a greater number of paid bookings. This was theorized because individual's closer to downtown possibly has easier access to music opportunities.

Visualization 8 depicts the relationship between annual spending on publicity (dollars) and distance from downtown (miles). The purpose of this visualization was to connect our first research question about spending to the second question that centers on distance from downtown (miles). This visualization shows little to no correlation between the two variables - with a small, negative correlation coefficient of -0.05. Seeing how spending was proven to play a role in number of paid bookings, this nonexistent trend between spending and distance from downtown further suggests that distance from downtown doesn't play a role in number of paid bookings artists obtain. This was surprising, because we had hypothesized that distance from downtown would be a significant predictor of spending and number of paid bookings in the end.

Some ethical implications include predicting a musician's success with bookings based on solely spending and income. This could potentially raise issues by discriminating against individual's who don't have enough money or resources to succeed (economic disparities). This firstly may not be fully representative of factors that can lead to success and secondly could be seen as labelling people with insufficient funds as "unable to succeed" in the music industry.

5) Reflection

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.

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 (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)