ECON 145: Final 1 Writeup

Robin Hollingsworth (PERMID: 3010287)

Hello Festival Party Planner!

I am writing to report back on the data collected from Boulder, Colorado's and their city's major festival. In comparison to last time, I have done a further analysis of all the variables and included data visualizations.

284 festival goers responded to a survey regarding information on travel and spending logistics surrounding the festival and their time in Boulder. Personal data for the visitors was also collected, such as age, gender, and home zip code.

When cleaning the data, I removed all duplicate data to accurately represent the sampled group. In regards to the request for using the upper bounds for the data that was provided with an interval, I do not believe that this is a fair way to present the data. Using upper bounds could lead to falsely high numbers. In this situation, it is more reliable and ethical to use the middle of the range provided for attributes such as age or amount spent. This way we can more accurate analyze the results of these spending amounts and properly form opinions about the success of the festival. Therefore, when reporting the amount spent on food, shopping, lodging, transportation, etc., I have included the average amount spent calculated using the lowest value, highest value, and mean value of the given range.

```
f1_dat <- read_csv("final1.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   X = col_double(),
##   hours_attend = col_double(),
##   zip = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
f1_unique <- distinct(f1_dat)

cleaned_min <- function(x){
  x <- gsub("Prefer*",NA,x)
  x <- gsub("\\-.*", "", x)
  x <- gsub("\\,", "", x)
  gsub("\\$","", x) %>% as.numeric()
}
cleaned_max <- function(x){
  x <- gsub("Prefer*",NA,x)
  x <- gsub(".*\\-", "", x)
  x <- gsub(" [A-z]*", "", x)
  x <- gsub("\\,", "", x)
  gsub("\\$","", x) %>% as.numeric()
}
cleaned_mid <- function(x){
  low <- cleaned_min(x)
  high <- cleaned_max(x)
  (high + low) / 2
}
cleaned_col <- function(data_vector, output_type){
  if (output_type == "min") {
    col <- cleaned_min(data_vector)
  }
```

```
  else if (output_type == "mid") {
    col <- cleaned_mid(data_vector)
  }
  else {
    col <- cleaned_max(data_vector)
  }
  return(col)
}
columns <- list(f1_unique$spend_food_drink,
                f1_unique$spend_private,
                f1_unique$spend_clothes,
                f1_unique$spend_transportation,
                f1_unique$spend_donations,
                f1_unique$spend_other,
                f1_unique$spend_food_drink_total,
                f1_unique$spend_entertainment_total,
                f1_unique$spend_shopping_total,
                f1_unique$spend_travel_total)
mean_min <- c()
mean_mid <- c()
mean_max <- c()
for (col in columns){
  mean_min %<>% append(col %>% cleaned_col('min') %>% mean(na.rm=TRUE))
  mean_mid %<>% append(col %>% cleaned_col('mid') %>% mean(na.rm=TRUE))
  mean_max %<>% append(col %>% cleaned_col('max') %>% mean(na.rm=TRUE))
}
variable <- c("spend_food_drink",
              "spend_private",
              "spend_clothes",
              "spend_transportation",
              "spend_donations",
              "spend_other",
              "spend_food_drink_total",
              "spend_entertainment_total",
              "spend_shopping_total",
              "spend_travel_total")
avgs <- tibble(variable, mean_min, mean_mid, mean_max)
```

Demographics

```
f1_unique %<>% mutate(local = ifelse(zip %in% c(80305, 80309, 80302, 80304, 80303,80306),1,0))
mean(f1_unique$local, na.rm=TRUE)
```

```
## [1] 0.8028169
```

```
mean(f1_unique$gender=="Female", na.rm=TRUE)
```

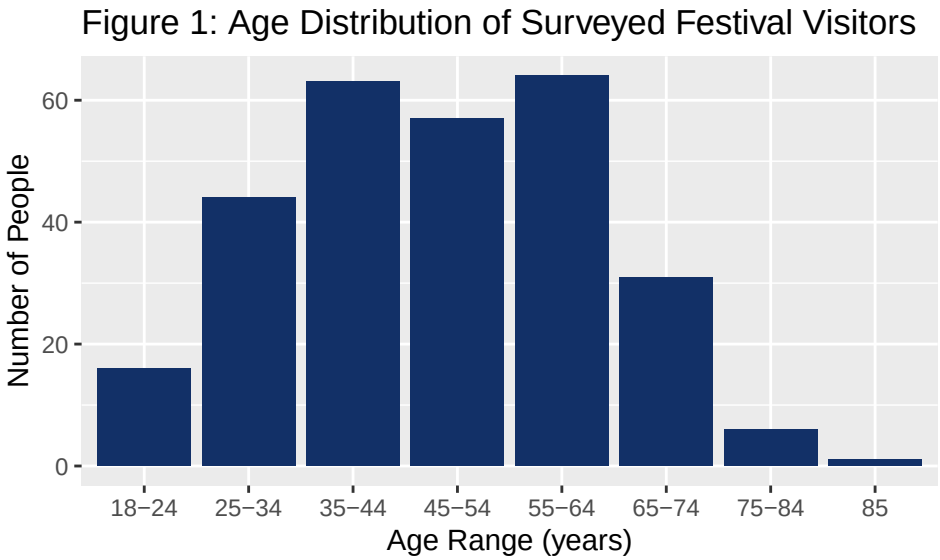
```
## [1] 0.6961131
```

Similar to my previous analysis, I looked at the most common zip codes of the people that were surveyed. Local zip codes (80305, 80309, 80302, 80304, 80303, 80306) made up **80.3%** of the surveyed festival goers. This is very high considering the main purpose of this festival was to generate national attention and attract visitors from outside Colorado to bring in revenue for the city of Boulder.

There was also a higher percentage of women surveryed at the festival than there were men. **69.6%** of the surveryed population was female and **30.4%** were male. This could mean that the festival was advertised or was catered more towards women than men.

```
ages <- f1_unique %>%
  mutate(age = ifelse(str_detect(age,"Prefer"), NA, age)) %>%
  filter(age != is.na(age))
```

```
ages %>%
  ggplot() +
  geom_bar(aes(age), fill="#123067") +
  ggtitle("Figure 1: Age Distribution of Surveyed Festival Visitors") +
  xlab("Age Range (years)") +
  ylab("Number of People")
```



As shown above in *Figure 1*, the most common age ranges are around 35-64 years old. There are very few attendees below the age of 24 and even less above the age of 75.

Stay and Travel Analysis

```
visit <- f1_unique$visits %>% cleaned_col('mid')
mean(visit > 1, na.rm=TRUE)
```

```
## [1] 0.9434629
```

```
days <- f1_unique$days_attend %>% cleaned_col('mid')
mean(days, na.rm=TRUE)
```

```
## [1] 2.985915
```

```
mean(f1_unique$hours_attend, na.rm=TRUE)
```

```
## [1] 3.581851
```

```
mode <- f1_unique %>% filter(!is.na(length_boulder)) %>% group_by(length_boulder) %>% summarize(count=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
mode
```

```
## # A tibble: 11 x 2
##    length_boulder     count
##    <chr>              <int>
##  1 Two nights            32
##  2 Day trip              30
##  3 Four nights           28
##  4 Local                 26
##  5 Resident              26
##  6 Local resident        25
##  7 Multiple day trips    22
##  8 One night             22
##  9 Three nights          21
```

```
## 10 I live here           17
## 11 Live in Boulder       17
```

Here I analyzed some of the factors contributing to the festival attendees' stay and their visitation of the festival. Interestingly, over **69%** of people had visited the festival more than 5 times, which means that the festival has a good return rate and people want to come back. However, only **5%** of people were first time visitors. This could have an impact on the growth of the festival.

For the people that responded to the survey, the average number of days attending the festival was **2.99 days** and the average number of hours spent at the festival was **3.58 hours**. The average number of days attending and the average visits were both calculated using the middle of the range that was given to neautralize over- or underestimations. The length of trips to Boulder for festival attendees who were outside visitors(not local residents/lived in Boulder) is shown in *Figure 2*.

Figure 3 separates the various types of lodging that was used by festival visitors. Only **25.3%** paid for a hotel/motel while staying in Boulder, contributing to the generated revenue expected from the festival.

```
length <- f1_unique %>% filter(length_boulder != is.na(length_boulder)) %>%
  filter(!(length_boulder  %in% c("I live here","Live in Boulder","Local","Local resident","Resident")))%>%
  group_by(length_boulder) %>%
  summarize(count=n()) %>%
  mutate(perc = round(100 * count/sum(count), digits=1), newLabels = str_c(length_boulder, "\n", perc, "%"
  arrange(desc(count))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
pie(length$count,
    labels=length$newLabels,
    main="Figure 2:\nLength of Boulder Stay for Non-Residents",
    col=c("#5da8b7","#b2b2b2","#d47c16","#7e4848", "darkgreen","#123067")
    )

lodge <- f1_unique %>% filter(lodging != is.na(lodging)) %>% group_by(lodging) %>% summarize(count=n()) %>%
  mutate(perc = round(100 * count/sum(count), digits=1), newLabels = str_c(lodging, "\n", perc, "%"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
pie(lodge$count,
    labels=lodge$newLabels,
    main="Figure 3:\nLodging Type Utilized While Attending Festival",
    col=c("#123067", "#5da8b7","#b2b2b2","#d47c16")
    )
```

Figure 2:
Length of Boulder Stay for Non−Residents

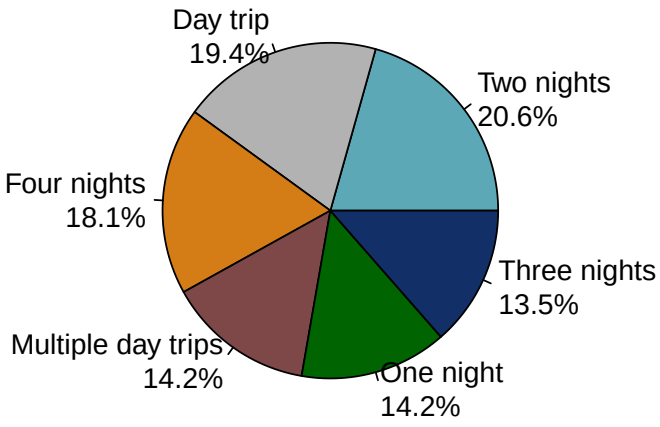
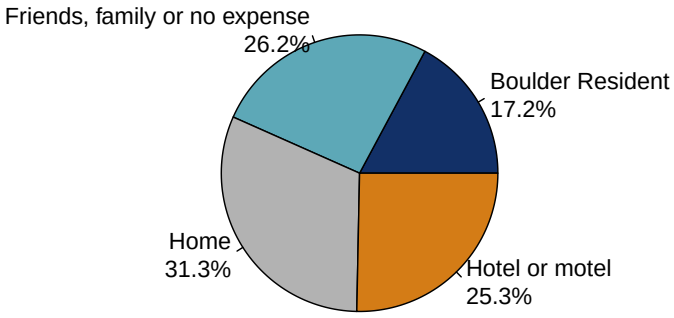


Figure 3:
Lodging Type Utilized While Attending Festival



Spending Analysis

Lastly, I analyzed the spending patterns of the surveryed festival goers. I started by calulating the average amount of money spent in each of the categories, seen in *Table 1*. To avoid sharing misleading data, I provided averages calculated using the minimum, middle, and maximum of the intervals inorder to provide the truest picture of the data.

```
kable(avgs, caption="Spending Averages")
```

Table 1: Spending Averages

variable	mean_min	mean_mid	mean_max
spend_food_drink	64.04270	98.69929	133.3559
spend_private	129.90975	176.26715	222.6245
spend_clothes	92.70652	124.48913	156.2717
spend_transportation	108.85663	150.51075	192.1649
spend_donations	88.89299	127.59963	166.3063
spend_other	384.66791	437.65485	490.6418
spend_food_drink_total	384.94141	645.38867	905.8359
spend_entertainment_total	103.30970	161.54291	219.7761
spend_shopping_total	131.54118	205.52745	279.5137
spend_travel_total	119.16863	188.00784	256.8471

The two categories where people tend to spend the most on are total food and drink and other. Although we aren't sure what other entails, food is purchased by both residents and non-resident, female and male, and young and old. So, it makes sense that it would have the highest total money spent as it is something that appeals to all.

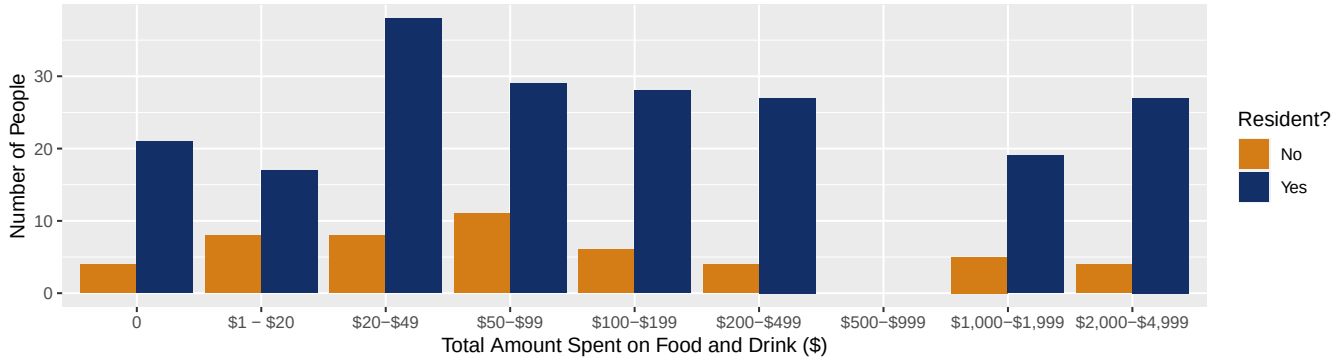
From this table we can also see major difference between mean_min and mean_max. This proves that using only mean_min or only mean_max can have an effect on our analysis of the data. For example, the difference between mean_min and mean_max for *spend_donations* is almost double.

In *Figure 4*, we can see which spending interval most people fell into for the category with the highest spending averages: *spend_food_drink_total*. In total, most people spent between $20-$49 on food and beverages while they were in Boulder. When separating the data into residents vs non-residents, we can see differences between the two groups. Overall, we can see that majority of the people who bought food were locals and the total number of non-residents that spent money on food and drink was much lower. More interestingly, for non-residents, the category with the most vistors was actually the $50-$99 category and for residents $20-$49 was the most popular. So, non-residents were spending a little bit more per person on food and drink but since there were so many more local residents, the locals' purchases are what make up majority of the money spent on food and drink in Boulder.

The reason the means calculated in *Table 1* are so much higher than the most common spending intervals is because of the handful of visitors (both local and non-local) that spent between $1,000-$5,000 on food and drink while in Boulder. These could be visitors that had longer stays or visitors that spent money on more expensive food options.

```
food <- f1_unique %>%
  mutate(spend_food_drink_total = ifelse(str_detect(spend_food_drink_total,"N/A"), NA, spend_food_drink_to
  filter(spend_food_drink_total != is.na(spend_food_drink_total)) %>%
  mutate(local = ifelse(zip %in% c(80305, 80309, 80302, 80304, 80303,80306),"Yes","No"))

food %>%
  ggplot() +
  geom_bar(aes(x=spend_food_drink_total, fill=local), position="dodge") +
  scale_x_discrete(limits = c("0","$1 - $20","$20-$49","$50-$99","$100-$199","$200-$499",
                              "$500-$999","$1,000-$1,999","$2,000-$4,999")) +
  labs(x="Total Amount Spent on Food and Drink ($)", y="Number of People") +
  scale_fill_manual(name="Resident?", values=c("Yes"="#123067", "No"="#d47c16")) +
  ggtitle("Figure 4: Total Dollars Spent on Food and Drink While in Boulder (Resident vs NonResident)")
```

Figure 4: Total Dollars Spent on Food and Drink While in Boulder (Resident vs NonResident)

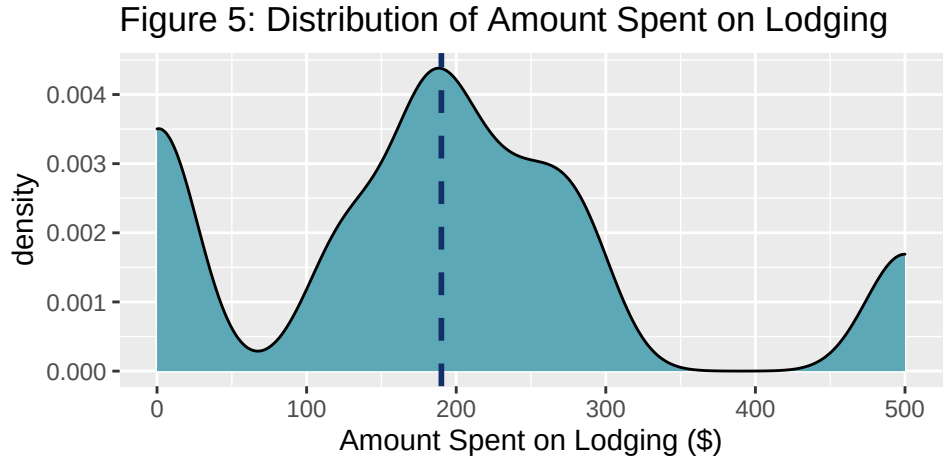



```
lodgeCost <- f1_unique %>%
  mutate(lodging_cost = ifelse(str_detect(lodging_cost,"N/A"), NA, lodging_cost)) %>%
  filter(!is.na(lodging_cost))
lodgeCost <- as.numeric(lodgeCost$lodging_cost)

mean(lodgeCost)
```

```
## [1] 190.0844
```

Lastly, *Figure 5* shows the distribution of the amount spent on lodging while visiting Boulder. The mean dollar amount spent was **$190.08** and is shown by the dashed dark blue line. As we can see the majority of visitors spent between $100-$300 on lodging while in Boulder. There was also a large amount of people that spent little to no money on lodging. This makes sense with the large amount of locals that were surveyed at the festival, given that they can easily stay at their own homes for no cost and still attend the festival. However, there were a select few visitors that chose to spend over $450 on lodging.

```
lodgeCost <- f1_unique %>%
  mutate(lodging_cost = ifelse(str_detect(lodging_cost,"N/A"), NA, lodging_cost)) %>%
  filter(!is.na(lodging_cost))
lodgeCost <- as.numeric(lodgeCost$lodging_cost)

tibble(lodgeCost) %>%
  ggplot() +
  geom_density(aes(x=lodgeCost), fill="#5da8b7") +
  geom_vline(xintercept=mean(lodgeCost), color="#123067", lty=2, size=1) +
  ggtitle("Figure 5: Distribution of Amount Spent on Lodging") +
  labs(x="Amount Spent on Lodging ($)")
```



Conclusion

Without more information on the cost to run the festival and other factors that go into it, it is hard to give a straight answer on whether or not the festival was a success. Overall, the data showed that the Boulder, Colorado Festival did attract a good number of attendees, but a lot of them were locals. With the goal of the festival being to bring in

revenue to the city and increases tourists, it struggled to bring in many long term, out-of-town visitors. The main demographic were middle aged people, mostly women. The area where visitors spent the most money was on food and drink. Below are tables and summary statistics for all of the variables for easy access. Hopefully, these insights from the data can help make the Boulder Festival more profitable and beneficial next year!

Summary Statistics

```
aColumns <- list(f1_unique$visits,
                 f1_unique$days_attend,
                 f1_unique$hours_attend,
                 f1_unique$age,
                 f1_unique$spend_food_drink,
                 f1_unique$spend_private,
                 f1_unique$spend_clothes,
                 f1_unique$spend_transportation,
                 f1_unique$spend_donations,
                 f1_unique$spend_other,
                 f1_unique$spend_food_drink_total,
                 f1_unique$spend_entertainment_total,
                 f1_unique$spend_shopping_total,
                 f1_unique$spend_travel_total)
variable <- c("visits",
              "days_attend",
              "hours_attend",
              "age",
              "spend_food_drink",
              "spend_private",
              "spend_clothes",
              "spend_transportation",
              "spend_donations",
              "spend_other",
              "spend_food_drink_total",
              "spend_entertainment_total",
              "spend_shopping_total",
              "spend_travel_total",
              "lodging_costs")

mean_min <- c()
mean_mid <- c()
mean_max <- c()
min_val <- c()
max_val <- c()

for (col in aColumns){
  mean_min %<>% append(col %>% cleaned_col('min') %>% mean(na.rm=TRUE))
  mean_mid %<>% append(col %>% cleaned_col('mid') %>% mean(na.rm=TRUE))
  mean_max %<>% append(col %>% cleaned_col('max') %>% mean(na.rm=TRUE))
  min_val %<>% append(col %>% cleaned_col('min') %>% min(na.rm=TRUE))
  max_val %<>% append(col %>% cleaned_col('max') %>% max(na.rm=TRUE))
}

mean_min %<>% append(mean(lodgeCost))
mean_mid %<>% append(mean(lodgeCost))
mean_max %<>% append(mean(lodgeCost))
min_val %<>% append(min(lodgeCost))
max_val %<>% append(max(lodgeCost))

allAvgs <- tibble(variable, mean_min, mean_mid, mean_max, min_val, max_val)
kable(allAvgs)
```

variable	mean_min	mean_mid	mean_max	min_val	max_val
visits	13.070671	13.266784	13.462898	0	56
days_attend	2.580986	2.985915	3.390845	0	40
hours_attend	3.581850	3.581850	3.581850	0	20
age	43.361702	47.760638	52.159574	18	85
spend_food_drink	64.042705	98.699288	133.355872	0	1000
spend_private	129.909747	176.267148	222.624549	0	3000
spend_clothes	92.706522	124.489130	156.271739	0	1499
spend_transportation	108.856631	150.510753	192.164875	0	3000
spend_donations	88.892989	127.599631	166.306273	0	1499
spend_other	384.667910	437.654851	490.641791	0	5000
spend_food_drink_total	384.941406	645.388672	905.835938	0	4999
spend_entertainment_total	103.309702	161.542910	219.776119	0	999
spend_shopping_total	131.541177	205.527451	279.513726	0	999
spend_travel_total	119.168627	188.007843	256.847059	0	999
lodging_costs	190.084388	190.084388	190.084388	0	500

Gender

- Female: 69.61%
- Male: 30.04%
- Prefer Not To Answer: 0.35%

Zip Code

- Most Common Zip Code: 80303
- Locals: 80.28%
- Non-Locals: 19.71%

Extend

- Extended Stay: 40.11%
- Did Not Extend Stay: 59.89%

Length (Most Common Responses)

1. Didn't extend stay (29)
2. Well, we try not to extend our stay, but the traffic blockage caused by the Festival slow us down. (29)
3. Live in area (28)

Length_Boulder (Most Common Responses)

1. Two Nights (32)
2. Day Trip (30)
3. Four Nights (28)

Lodging (Most Common Responses)

1. Boulder Resident (40)
2. Friends, family or no expense (61)
3. Home (73)
4. Hotel or motel (59)