Andres Cojuangco & Pedro Dutra

QAC 211

Final Report

Introduction:

For our final project, we investigated the Iowa Liquor Sales public dataset on BigQuery. This dataset contained liquor sales data dating back to 2012 all the way to present day. Information included, the date of the purchase, the location of where the product was sold, the location of the store, the zip code of the store, the volume of alcohol sold to each store, the retail price of each drink, among others. Immediately, we were drawn to investigating consumption habits of Iowans, making the only challenge finding what variables we should hold constant. After some consideration, we opted to investigate which zip codes were the biggest contributors to the sale of luxury liquor brands. In addition, we investigated how the top performing zip codes' consumption changed during the pandemic compared to the rest of Iowa's zip codes.

To collect Iowa's zip codes, we had to use the ESRI Tapestry dataset. The ESRI dataset had information on every American zip code, including lifestyle, median household income, and median age. Therefore, we were able to find more information on each zip code after merging the ESRI dataset with Iowa's liquor sales.

The hypothesis for this research project was that luxury brands would excel across Iowa. However, they will proportionally do better in urban areas due to the higher median income of the residents in the urban areas. If the hypothesis is correct, the results should indicate that the luxury liquor brands are consumed in more affluent zip codes and that other brands should perform best in lower-income neighborhoods in both urban and rural settings.

Data Management:

One of the biggest challenges of this project was cleaning the data frames. After merging the ESRI dataset to the Iowa Liquor Sales dataset by zip code, we figured out how to categorize certain variables to answer our research question.

The first variable we had to modify was the variable that categorized each zip code into urban, suburban and rural. Creating this variable was useful because it allows us to understand the zip codes with highest and lowest consumption better. This categorization was done using the Iowa tapestry data, from ESRI. We filtered through the variable "TURBZNAME" which gave a description of each zip code. Then, we paired all the urban, suburban and rural zip codes into different categories.

Categorizing luxury brands was also imperative for this project. Unfortunately, there was no variable that allowed us to speed up the categorization, therefore we had to use our knowledge of alcohol beverages to categorize luxury liquor brands. Thus, we went through the brands that we recognized as luxury liquor brands under the variable item_description and categorized them. The data was queried from BigQuery with a text search of the luxury liquor brands.

With those two new variables, we were able to create a new data frame, where we found the total sales of luxury brands grouped by year. However, we found the discrepancy between luxury brands and other brands to be extremely high in favor of the non-luxury liquor brands. This was problematic because in our hypothesis we expected a much more balanced distribution in sales, particularly in urban settings. Then, when finding the percentage of sales of luxury brands per zip code, we found that no zip codes exceeded 1% in terms of luxury brands bought

over total sales. Additionally, luxury brand sales were zero in some zip codes. The difference in sales for luxury brands and non-luxury brands required us to pivot our research. The new research questions were 1) which zip codes were the biggest contributors to the sales of luxury liquor brands and 2) how did the sales of luxury liquor brands in those zip codes compare to the rest of Iowa during the pandemic?

As a result, ***Figure 1*** and ***Figure 2*** were created. In ***Figure 1***, it is clear how across all years, non-luxury liquor brands excel while luxury brand sales are not remotely close. In ***Figure 2***, only the zip codes with at least one sale of luxury brands were included in the map. Using the function choropleth, the zip codes that contribute to the highest sale of luxury liquor brands were highlighted in red. These percentages were from 2017 as the ESRI Tapestry dataset from Iowa is from 2017 too.

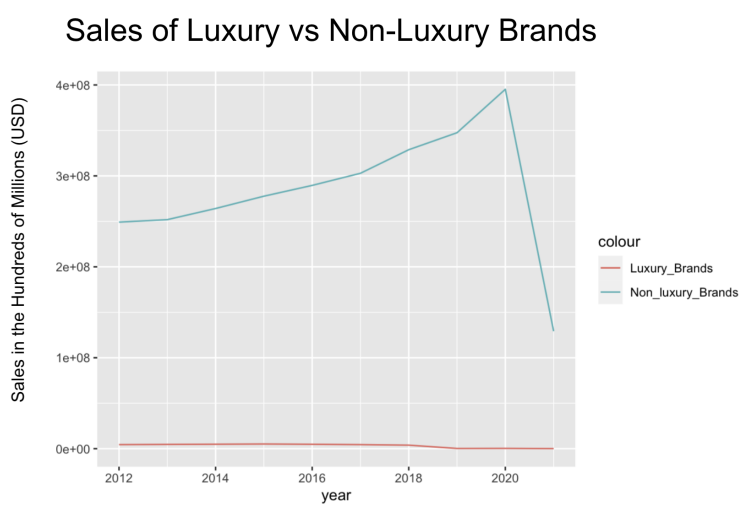

Figure 1

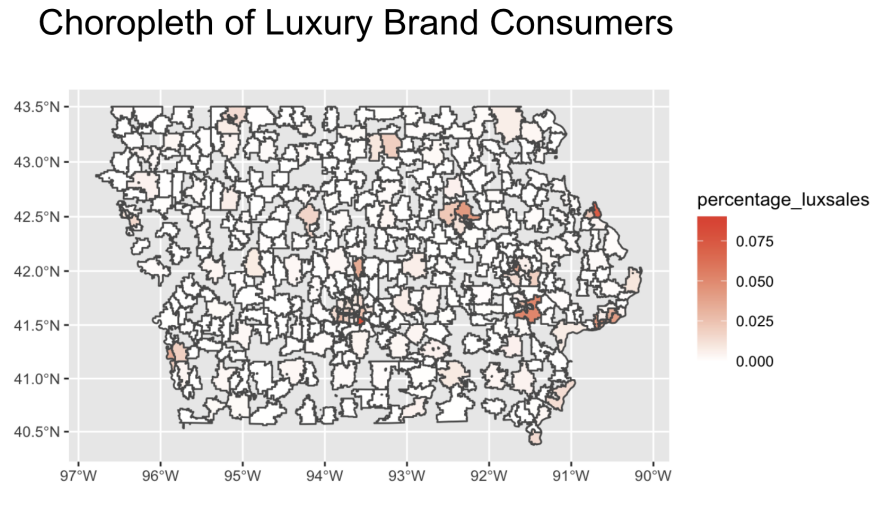

Figure 2

Top 6 Zip Codes:

Despite the consumption patterns being heavily skewed towards non-luxury brands, we still filtered the data to find the six zip codes that proportionally consumed the most luxury branded liquor. Answering the first new research question, the zip codes found were 50314, 50320, 52001, 52240, 52804, 50703. These zip codes represented the following towns, Des Moines (for the first two), Dubuque , Iowa City, Davenport, and Waterloo respectively.

All zip codes were categorized as urban except for 50320 which was categorized as suburban. Furthermore, all of the top six zip codes were under the median age and household income for Iowa. The ESRI Tapestry data for Iowa also allowed us to investigate the type of lifestyle in each zip code, and all of them were found to be lower income neighborhoods. This deviated from the hypothesis partially because while we predicted that the luxury brands would excel in urban settings, it was also assumed that they would excel among wealthier zip codes. Nevertheless, even with a higher price, the top consumers of luxury brands come from lower-income neighborhoods. This indicates that price may not be an important factor and there may be other motivators driving the sales of luxury liquor brands in these zip codes.

COVID-19 Time Series Data:

For the second new research question, we hypothesized that the luxury liquor brand sales would perform better in the top six zip codes compared to the rest of Iowa because they would be more affluent zip codes with more allowance to spend on luxury liquor despite the pandemic. But due to our speculation that price is not an important factor from our previous finding, we wanted to confirm this by observing the proportion of sales of luxury liquor brands in the top six zip codes and the proportion of sales of luxury liquor brands for the rest of Iowa (luxury liquor

brand sales in six zip codes / total sales of six zip codes & luxury liquor brand sales for the rest of Iowa / total sales for the rest of Iowa) over a time series by month. We filtered the time series data so that we would only get the values from January 2020 onwards, as we were interested in consumption habits during the lockdown period.

Before exploring the results, it is important to note that Iowa was quite lax in its approach regarding the COVID-19 pandemic. The state only closed schools at the end of March and a statewide lockdown was never officially declared by Governor Kim Reynolds. Additionally, a mask mandate was never put in place, as Gov. Reynolds stated that she believed "the people would do the right thing." This is important to note because while Iowa might have experienced some lockdown fatigue, it was probably not as severe as other states where the mandates were much stricter. Furthermore, bars remained open in some areas of Iowa throughout the pandemic, meaning people did not have to buy more from liquor stores, in theory.

After creating the variable for the percentage of luxury liquor brand sales per month for the two data frames, the results found that the top six zip codes had a higher percentage for 10 of the 16 months in the data frame. Despite that, the maximum point belongs to the other zip codes in Iowa in the month of October, as we can see on ***Figure 3***. It is important to note that Iowa did see a spike in cases around September. However, the peak in cases for COVID-19 in Iowa was in November, which corresponds to the peak in luxury brand sales for the top six zip codes. Nevertheless, the data in ***Figure 3*** are all extremely small, under 1%, but it still displays consumption habits for the top six zip codes in comparison to the rest of Iowa effectively. The fact that the proportion of sales of luxury liquor brands in the top six zip codes performed just as well as the proportion of sales of luxury liquor brands for the rest of Iowa further supports that price is not an important factor when purchasing these luxury brands. Additionally, it shows that

there is no real deviation for consumption patterns between the top six zip codes and the rest of Iowa during the pandemic.

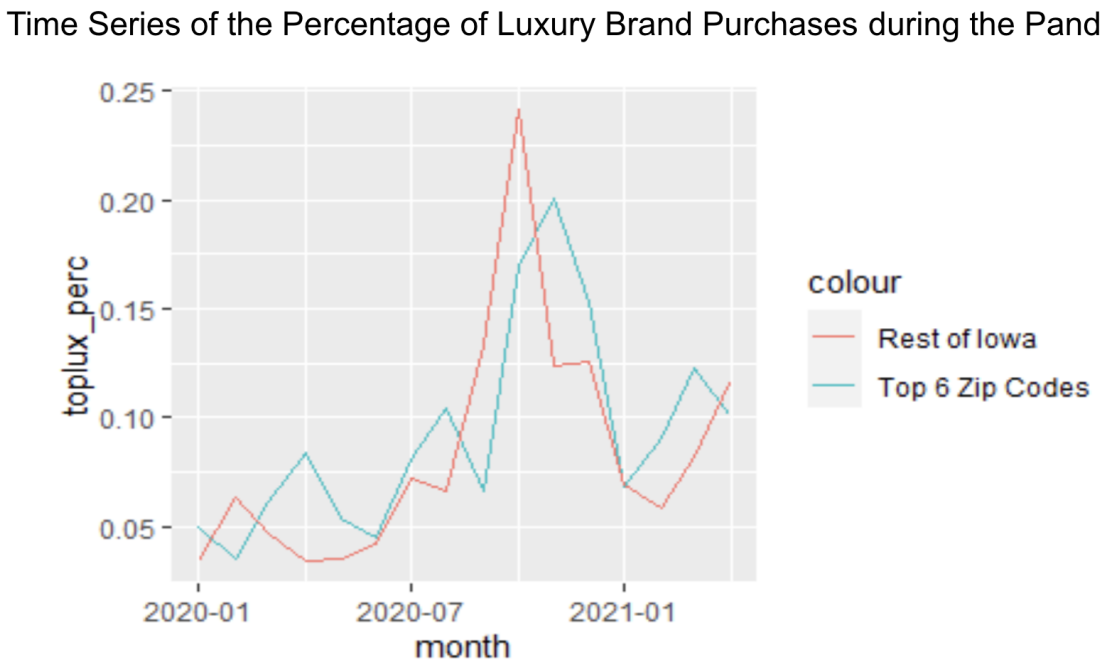

Figure 3

Further Research:

Despite our hypotheses being incorrect, the findings provide us information that can lead us in the right direction in further understanding the consumption patterns of Iowa. In continuation with this project, our new data frame can be used for many research purposes after some modifications, including market segmentation. While we opted to dive into the zip codes and luxury liquor brands, there is a lot more to dive into, such as census data, volumes sold,

stores sold, etc. Additionally, some regressions can be run on these patterns, if the year is factored into it. From there, we can dive deeper into the project by investigating the most popular brands or kinds of liquor across Iowa and by smaller categories. Not to mention, this project can be a step towards generating recommendations for liquor stores in Iowa such as projecting sales or predicting inventory.

Limitations:

One limitation is that the luxury brands were manually chosen from a list of all the liquor brands in the Iowa Liquor Sales dataset. Because we are not aware of all liquor brands in the data set, there may be more luxury brands that were left out. Human error may have been another factor if some brands were left out. One way we can improve this is by creating a clear definition of a luxury liquor brand and setting criteria for what makes a luxury brand.

Another small limitation was we could not run a regression analysis on the dataset. The reason for that is that once we collected the top six zip codes, we only had four degrees of freedom, meaning running a linear regression would not be the most efficient way to analyze the data. Additionally, we would have to subset by month which would make the analysis process more difficult. If the research question was pivoted, then perhaps a regression model could be created from this dataset.

Appendix:

##All Libraries

```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)
library(lubridate)
library(sf)
------
## Cleaning Data for ESRI Tapestry (2017) Data Set
load("iowa_tap_2017.RData", verbose=T)



#Creating Dummy Variables for Rural (0), Suburban (1) and Urban Settings (2)
ia_tapestry$zc_category[ia_tapestry$TURBZNAME == "Rural" | ia_tapestry$TURBZNAME ==
"Semirural"] <- 0

ia_tapestry$zc_category[ia_tapestry$TURBZNAME == "Suburban Periphery"] <- 1

ia_tapestry$zc_category[ia_tapestry$TURBZNAME == "Urban Periphery" |
ia_tapestry$TURBZNAME == "Metro Cities" | ia_tapestry$TURBZNAME == "Principal Urban
Center"] <- 2

names(ia_tapestry)[names(ia_tapestry) == "ID"] <- "zip_code"
```

```
#Filtering out columns from ia_tapestry that we don't need and joining it to bigbrands_df

main_df = ia_tapestry %>% select(NAME, zc_category, zip_code) %>%
inner_join(bigbrands_df, ia_tapestry, by = "zip_code")

main_df
-----

## Calculating the statewide percentage of sales for big and small brands

ttl_sales_per_year <- main_df %>% group_by(year, is_brand) %>% summarize(t_sales =
sum(sale_dollars)) %>% select(year, is_brand, t_sales) %>% sf::st_drop_geometry()

pivot_ttl_sales_per_year <- ttl_sales_per_year %>% pivot_wider(names_from = is_brand,
values_from = t_sales)

names(pivot_ttl_sales_per_year)[names(pivot_ttl_sales_per_year) == "0"] <-
"Non_luxury_Brands"
names(pivot_ttl_sales_per_year)[names(pivot_ttl_sales_per_year) == "1"] <- "Luxury_Brands"
-----
#Line Graph Luxury vs Non
pivot_ttl_sales_per_year
ggplot(pivot_ttl_sales_per_year, aes(x = year)) +  geom_line(aes(y = Luxury_Brands, color =
"Luxury_Brands")) +
  geom_line(aes(y = Non_luxury_Brands, color = "Non_luxury_Brands")) + expand_limits(y = 0)
----
```

```
##calculating total sales for each zip code for 2017
t_sales_zip <- main_df %>% filter(year == 2017) %>% group_by(zip_code) %>%
summarize(tsales_zip = sum(sale_dollars)) %>% sf::st_drop_geometry()

df_filter_2017 <- main_df %>% filter(year == 2017) %>% sf::st_drop_geometry()

df_2017 <- merge(t_sales_zip, df_filter_2017, by.x = "zip_code", by.y = "zip_code")

t_sales_zip
df_filter_2017
df_2017
----
#Total Sales for 2017

ttl_sales_2017 <- ttl_sales_per_year %>% filter(year == 2017)
ttl_sales_2017_val <- sum(ttl_sales_2017$t_sales)

#Rearranging columns for df_2017

col_order <- c("NAME", "zc_category", "zip_code", "year", "is_brand", "volume_sold",
"sale_dollars", "tsales_zip")

df_2017 <- df_2017[,col_order]

#Rows with zip codes that only appear once and mutating the items that only appear once to
make is_brand == 1 sales
```

```
once <- subset(df_2017, ave(NAME,NAME,FUN=length) == 1) %>% mutate(is_brand = 1,
volume_sold = 0, sale_dollars = 0)

#Adding the missing rows (once) to the ttl_sales_2017

df_2017_combined <- rbind(df_2017,once)
df_2017 <- df_2017_combined %>% arrange(zip_code)

#Making df for choropleth and adding column for percentage of total sales for luxury and small
brands and filtering out small brands

choro_2017_1 <- df_2017 %>% filter(is_brand == 1) %>% mutate(percentage_luxsales =
(sale_dollars/ttl_sales_2017_val)*100)

top_lux_buyers <- arrange(choro_2017_1, desc(percentage_luxsales)) %>% select("NAME",
"zc_category", "zip_code", "sale_dollars", "percentage_luxsales")
head(top_lux_buyers)
----
##Making the Choropleth for the percentage of total sales spent on luxury brands

ia2 = ia_tapestry %>% inner_join(choro_2017_1, by="zip_code")

ggplot() + geom_sf(data = ia2, aes(fill=percentage_luxsales)) +
  scale_fill_gradient2(low = 'blue', high='red', mid='white')
```

```
#Import Time Series from BigQuery
tscovid_df <- read_csv("tscovid_df.csv")
View(tscovid_df)



#Create dates and group by month
tscovid_df$date = as.Date(with(tscovid_df, paste(year, month, day, sep="-")), "%Y-%m-%d")

ts20_21 = tscovid_df %>% filter(date >= as.Date("2020-01-01")) %>% group_by(month =
lubridate::floor_date(date, "month")) %>%  arrange(zip_code)
ts20_21

#Data frames for non-top zip codes
rest_of_iowa = ts20_21 %>% filter(zip_code != 50314 & zip_code != 50320 & zip_code != 52001
& zip_code != 52240 & zip_code != 52804 & zip_code != 50703)

sum_for_rest = rest_of_iowa %>% summarize(sales_month = sum(sale_dollars))

sum_lux_rest = rest_of_iowa %>% filter(is_brand==1) %>% summarize(sales_month =
sum(sale_dollars))

rest = inner_join(sum_for_rest, sum_lux_rest, by="month", suffix = c("total_sales",
"luxury_sales"))

names(rest)[names(rest) == "sales_monthtotal_sales"] <- "total_sales"
```

```
names(rest)[names(rest) == "sales_monthluxury_sales"] <- "luxury_sales"

rest



#Data frames for top zip codes in Iowa
sum_for_top = topzip %>% summarize(sales_month = sum(sale_dollars))

sum_top_one = topzip %>% filter(is_brand==1) %>% summarize(sales_month =
sum(sale_dollars))

ratio_top = inner_join(sum_for_top, sum_top_one, by="month", suffix = c("total_sales",
"luxury_sales"))

names(ratio_top)[names(ratio_top) == "sales_monthtotal_sales"] <- "top_total_sales"

names(ratio_top)[names(ratio_top) == "sales_monthluxury_sales"] <- "top_luxury_sales"

ratio_top

#Creating Proportions
ratio_top = ratio_top %>% mutate(toplux_perc = 100* (top_luxury_sales / top_total_sales))

rest = rest %>% mutate(notlux_perc = 100 * (luxury_sales / total_sales))
```

```
#Selecting Variables to Join

x = ratio_top %>% select(month, toplux_perc)

y = rest %>% select(month, notlux_perc)


rest_vs_top = inner_join(x, y, by="month")

rest_vs_top

#Time series graph

ggplot(rest_vs_top, aes(x = month)) + geom_line(aes(y = toplux_perc, color = "Top 6 Zip

Codes")) + geom_line(aes(y = notlux_perc, color = "Rest of Iowa"))
```