Data Preparation

DATA 403: Project 1

Justin Mai, Libby Brill, Maxwell Dubow, Rachel Hartfelder

Data Collection

- Added 2 sources of outside data
  - https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html
    - Census population estimate, most accurate up-to-date data, stated as best practice to use this data for population totals by Census.gov.
  - https://en.wikipedia.org/wiki/List_of_colleges_and_universities_in_Iowa
    - List of Universities in Iowa and their corresponding city
- We chose to focus on the last 5 years as we wanted higher velocity data, for our predictive and interpretive models.
  - We used the last year for initial overview and to set up the process, and will be expanding to the last 5 years once our model setup is complete

Data Cleaning

- Removing rows with negative "Bottles Sold"
- Removed Columns
  - We removed `Invoice/Item Number`, `Store Number`, `Store Name`, `Address Zip Code`, `Store Location`, `County Number`, `Category`, `Vendor Number`, `Vendor Name`, `Item Number`, `Item Description`, `Pack`, `Bottle Volume (ml)`, `State Bottle Cost`, `State Bottle Retail`, `Bottles Sold`, `Volume Sold (Gallons)`
    - Since we don't care about store location, what stores, store name, and by association zip code, along with that specific invoice along with county number so those were removed. We also choose to remove categories since we already have category names. Additionally, we also choose to remove the Item number as we have no reference to that value. Item description was removed as Category

names exist. Pack, state bottle cost, state bottle retail, bottle volume, volume sold, and bottles sold as we have total sales for that specific item of alcohol and we kept volume sold (liters) as our metric of measurement.

Feature Engineering

- Added Columns
  - Added a column for colleges in Iowa corresponding to cities within Iowa.
  - `CollegeTF` column of boolean values, true if city has college, false if no.
  - Added a `PopYear` column matching with the year given for the census data
  - Condensed the types of Alcohol into 10 distinct categories with a column called categories, Whiskey, Vodka, Rum, Gin, Tequila/Mezcal, Brandy, Cordials & Liqueur, Ready to Drink, Neutral Grain Spirit
    - Used ChatGPT to classify types of alcohol
- Modified/Renamed Columns
  - From the Date column added a column called Year.
  - Transformed and renamed the Date column to Month.
  - Renamed `Category Name` to `Category`, `Sale (Dollars)` to `Sales`, `Volume Sold (Liters)` to `LitersSold`

Data Transformation

- Pivoted data from wide to long (around 3 main columns: `City`, `College`, and `CollegeTF`).
- Left with one large dataset called `df_iowa` with columns:
  - `City`: City in Iowa
  - `College`: College(s) belonging to city in Iowa
  - `CollegeTF`: True if College in city, False otherwise.
  - `PopYear`: Year of population estimate by U.S. Census
  - `Population`: Population estimate for city of given year by U.S. Census
  - `Month`: Month of year from original date column of Iowa Liquor Sales
  - `County`: County of city.
  - `Category`: One of 10 types of classification for Alcohol
  - `Sales`: Total cost of liquor order (number of bottles multiplied by the state bottle retail)
  - `LitersSold`: Total volume of liquor ordered in liters. (i.e. (Bottle Volume (ml) x Bottles Sold)/1,000)
  - `Year`: Year from original date column of Iowa Liquor Sales

Client A: `df_A`

- Grouped by `City`, `CollegeTF`, `Population`, `Category`, `Year`, `Month`
- Produced `AvgSales` by summing the city sales and dividing by the city population, for average sales per person (i.e. .43: .43 bottles sold per person in that city for that month within that year)

Client B: `df_B`

- Grouped by `City`, `CollegeTF`, `Population`, `Category`
- Produced `AvgLiters` by summing the city liters sold and dividing by the city population for average liters sold per person. (i.e. 1.6: 1.6 liters sold per person in that city for that month within that year)

Code

```
iowa_24_25 <- read_csv(here::here("Data", "Iowa_Liquor_Sales_2024_2025.csv"))
iowa_pop <- read_excel(here::here("Data", "Iowa_City_Populations.xlsx"),
                       skip = 3)
```

```
iowa_24_25 <- iowa_24_25 |>
  select(-c(`Invoice/Item Number`, `Store Number`, `Store Name`, `Address`,
            `Zip Code`, `Store Location`, `County Number`, Category,
            `Vendor Number`, `Vendor Name`, `Item Number`, `Item Description`,
            Pack, `Bottle Volume (ml)`, `State Bottle Cost`,
            `State Bottle Retail`, `Bottles Sold`, `Volume Sold (Gallons)`)) |>
  mutate(Date = mdy(Date),
         PopYear = case_when(Date < mdy("9/1/2021") ~ 2020,
                             Date < mdy("9/1/2022") ~ 2021,
                             Date < mdy("9/1/2023") ~ 2022,
                             Date < mdy("9/1/2024") ~ 2023,
                             TRUE ~ 2024),
         Year = year(Date),
         Date = month(Date, label = TRUE, abbr = TRUE),
         across(City:`Category Name`, str_to_title)) |>
  rename(Month = Date,
         Category = `Category Name`,
         Sales = `Sale (Dollars)`,
         LitersSold = `Volume Sold (Liters)`)
```

```r
df_iowa <- iowa_pop |>
  mutate(City = str_to_title(str_remove(`...1`, " city, Iowa$")),
         College = case_when(
           City == "Iowa City" ~ "University of Iowa",
           City == "Ames" ~ "Iowa State University",
           City == "Cedar Falls" ~ "University of Northern Iowa",
           City == "Ankeny" ~ "Des Moines Area Community College",
           City == "Davenport" ~ "Eastern Iowa Community Colleges;
           Palmer College of Chiropractic; St. Ambrose University",
           City == "Iowa Falls" ~ "Ellsworth Community College",
           City == "Waterloo" ~ "Hawkeye Community College; Allen College",
           City == "Ottumwa" ~ "Indian Hills Community College",
           City == "Fort Dodge" ~ "Iowa Central Community College",
           City == "Estherville" ~ "Iowa Lakes Community College",
           City == "Council Bluffs" ~ "Iowa Western Community College",
           City == "Cedar Rapids" ~ "Kirkwood Community College; Coe College;
           Mount Mercy University",
           City == "Marshalltown" ~ "Marshalltown Community College",
           City == "Mason City" ~ "North Iowa Area Community College",
           City == "Calmar" ~ "Northeast Iowa Community College",
           City == "Sheldon" ~ "Northwest Iowa Community College",
           City == "West Burlington" ~ "Southeastern Community College",
           City == "Creston" ~ "Southwestern Community College",
           City == "Sioux City" ~ "Western Iowa Tech Community College;
           Briar Cliff University; Morningside University;
           St. Luke's College",
           City == "Sioux Center" ~ "Dordt University",
           City == "Storm Lake" ~ "Buena Vista University",
           City == "Pella" ~ "Central College",
           City == "Dubuque" ~ "Clarke University; Emmaus University;
           Loras College; University of Dubuque;
           Wartburg Theological Seminary",
           City == "Epworth" ~ "Divine Word College",
           City == "Des Moines" ~ "Drake University; Des Moines University;
           Mercy College of Health Sciences; Grand View University",
           City == "Mount Vernon" ~ "Cornell College",
           City == "Lamoni" ~ "Graceland University",
           City == "Grinnell" ~ "Grinnell College",
           City == "Decorah" ~ "Luther College",
           City == "Fairfield" ~ "Maharishi International University",
           City == "Indianola" ~ "Simpson College",
           City == "Orange City" ~ "Northwestern College",
```

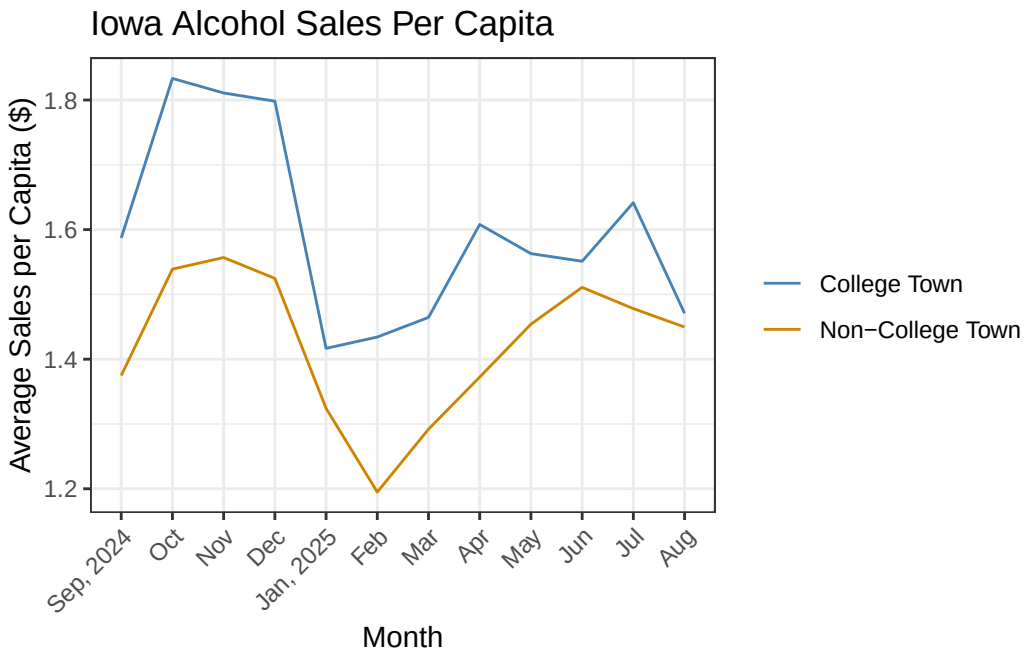
```
           City == "Fayette" ~ "Upper Iowa University",
           City == "Waverly" ~ "Wartburg College",
           City == "Oskaloosa" ~ "William Penn University",
           City == "Forest City" ~ "Waldorf University",
           TRUE ~ NA_character_),
         CollegeTF = !is.na(College)) |>
  select(City, `2020`, `2021`, `2022`, `2023`, `2024`, College, CollegeTF)

df_iowa <- df_iowa |>
  pivot_longer(cols = -c(City, College, CollegeTF),
    names_to = "PopYear",
    values_to = "Population"
  ) |>
  mutate(PopYear = as.numeric(PopYear)) |>
  inner_join(iowa_24_25, by = c("City", "PopYear")) |>
  filter(Sales > 0,
         LitersSold > 0) |>
  mutate(Category = case_when(
      str_detect(Category, regex("Whisk", ignore_case = TRUE)) ~ "Whisky",
      str_detect(Category, regex("Vodka", ignore_case = TRUE)) ~ "Vodka",
      str_detect(Category, regex("Rum", ignore_case = TRUE)) ~ "Rum",
      str_detect(Category, regex("Gin", ignore_case = TRUE)) ~ "Gin",
      str_detect(Category, regex("Tequila|Mezcal", ignore_case = TRUE))
      ~ "Tequila/Mezcal",
      str_detect(Category, regex("Brandi", ignore_case = TRUE)) ~ "Brandy",
      str_detect(Category, regex("Cordials|Liqueur|Schnapps|Triple Sec|
                                 Cream Liqueur|Coffee Liqueur",
                                 ignore_case = TRUE)) ~ "Cordials & Liqueur",
      str_detect(Category, regex("Cocktails|Rtd", ignore_case = TRUE))
      ~ "Ready to Drink",
      str_detect(Category, regex("Neutral Grain Spirits", ignore_case = TRUE))
      ~ "Neutral Grain Spirit",
      TRUE ~ "Other"))
```

```
df_A <- df_iowa |>
  group_by(City, CollegeTF, Population, Category, Year, Month) |>
  summarize(TotalSales = sum(Sales), .groups = "drop") |>
  mutate(AvgSales = TotalSales/Population) |>
  select(-TotalSales)
```

```r
df_B <- df_iowa |>
  group_by(City, CollegeTF, Population, Category) |>
  summarize(TotalLiters = sum(LitersSold), .groups = "drop") |>
  mutate(AvgLiters = TotalLiters/Population) |>
  select(-TotalLiters)
```

Data Visualizations

```r
df_A |>
  mutate(Month = factor(if_else(Month == "Sep", "Sep, 2024",
                                if_else(Month == "Jan", "Jan, 2025", Month)),
                        levels = c("Sep, 2024", "Oct", "Nov", "Dec",
                                   "Jan, 2025", "Feb", "Mar", "Apr", "May",
                                   "Jun", "Jul", "Aug")),
         CollegeTF = if_else(CollegeTF, "College Town",
                                    "Non-College Town")) |>
  group_by(CollegeTF, Month) |>
  summarize(AvgSales = mean(AvgSales), .groups = "drop") |>
  ggplot(aes(x = Month, y = AvgSales, color = CollegeTF, group = CollegeTF)) +
  geom_line() +
  scale_color_manual(values = c("steelblue", "orange3")) +
  labs(x = "Month", y = "Average Sales per Capita ($)",
       color = "", title = "Iowa Alcohol Sales Per Capita") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
weighted_by_cat <- df_B |>
  group_by(Category, CollegeTF) |>
  summarise(
    weighted_mean = weighted.mean(AvgLiters, w = Population, na.rm = TRUE),
    .groups = "drop"
  ) |>
  mutate(
    CollegeTF = factor(if_else(CollegeTF, "College Town", "Non-College Town"),
                       levels = c("Non-College Town", "College Town"))
  )

ggplot(weighted_by_cat, aes(x = Category, y = weighted_mean,
                            fill = CollegeTF)) +
  geom_col(position = position_dodge(width = 0.7), width = 0.65) +
  coord_flip() +
  scale_fill_manual(
    values = c("College Town" = "steelblue", "Non-College Town" = "orange3"),
    breaks = c("College Town", "Non-College Town"),
    labels = c("College Town", "Non-College Town")
  ) +
  labs(x = NULL, y = "Population-Weighted Average Liters per Person",
       fill = NULL) +
  theme_minimal(base_size = 11) +
```

```
  theme(panel.grid.minor = element_blank())
```