Data Mining Project Write-Up

Cristian De Leon, Talia Quiroga, Brianna Jacome

Executive Summary

In this project, we began with a dataset from Kaggle that focused on the amount of plastic waste produced by each country in 2023. Recognizing the need to broaden our analysis, we carefully selected four additional datasets from Gapminder and the World Bank Group to enhance the variety and depth of possible queries. This process required a thoughtful and collaborative approach, as we debated which datasets would be both interesting and insightful to analyze. Our intention was to focus on insights that were not only economically relevant but also socially beneficial. This posed a significant challenge, as many potential datasets had to be discarded due to our key defining criterion of all data needing to pertain to the year 2023. We prioritized using recent data because pollution and plastic waste are rapidly evolving issues, and outdated information might not accurately reflect current trends.

The original Kaggle dataset provided a solid starting foundation with five variables: main sources of plastic waste, recycling rates, total plastic waste (in metric tons), per capita plastic waste, coastal waste risk. To complement this, we chose additional datasets that covered: child mortality by country (ages 0–5), GDP per capita, population density, and urban population. These datasets allowed us to explore a wide range of questions, from economic disparities to environmental and social impacts through distribution analysis, correlation analysis, and relationship analysis. This deliberate combination ensured our analysis was both robust and meaningful, shedding light on the complexities of plastic waste and its global implications.

Cleaning the Data

The initial dataset about global plastic had a total of 165 observations, with each observation being a different row of data. First checking that there were no null fields in the data set that need to be removed, which luckily there was none. This was the same for all the datasets that we pulled. Then we proceeded to parse the GDP per capita data, population density, and child mortality data into individual tables only displaying the data from 2023, because these dataset had data from multiple years. This was done simplye by creating a separate table for each using the select() function and saving them as new dataframes.

Another issue we ran into was that the country names between the tables and the plastic waste dataset were different, with some country names being abbreviated. In order to merge the two datasets we needed to ensure that they had the same country values, in order to avoid losing data from mismatching country names. This was accomplished by using the mutate() function and the str_replace(), to make the abbreviations match the names in the Plastic waste dataset in each table.

Merging the data frames and the original dataset was done using the inner join function, which reduced the initial 165 observations to 146 observations. We also chose to add the continents category to the data as a way to aggregate data, which was by using the "countrycode" package in R, and merging it with the larger dataset.

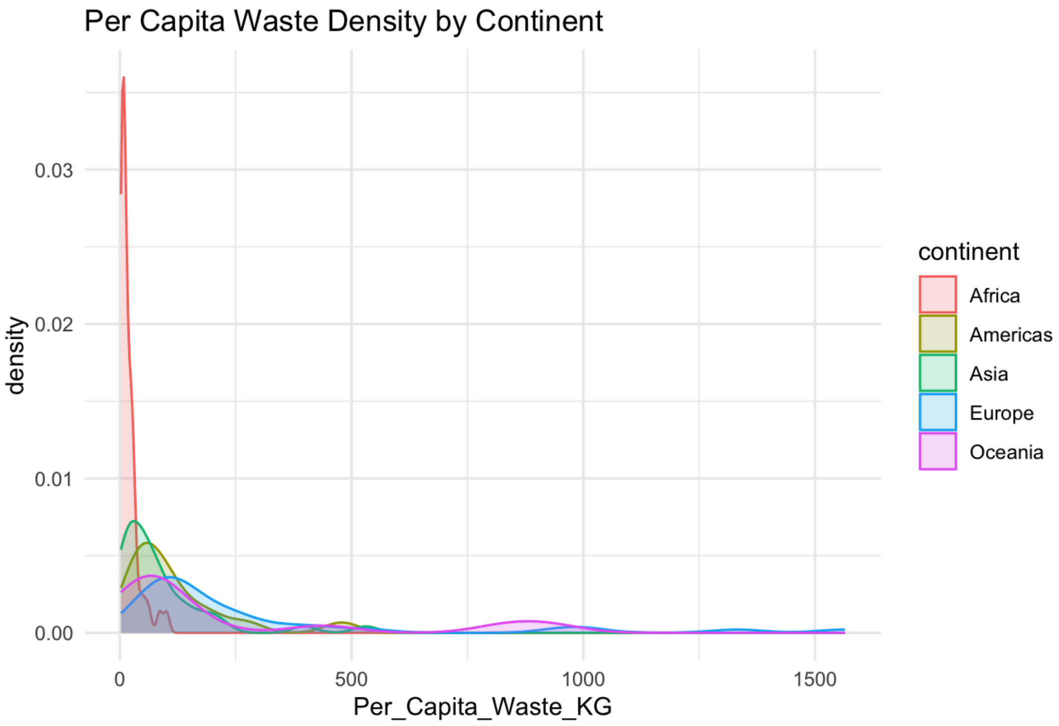
Once merged we ensured that all data was uniform and able to be used. The GDP data was being recognized initially as characters because the data was represented in thousands, and some values had a "K" next to the numerical value to represent that. Using the as.numeric function and the gsub function, we were able to multiply all values with a K at the end with 1000 to create a numeric value that matched the rest of the data. We also chose to convert the total plastic waste amount from metric tons to kilograms to match the units of the per capita waste production to be able to make any calculations needed and to have uniformity. This was done by simply multiplying the metric tons data by 1000. Finally we noticed that the category labels in plastic sources were in some cases simply flipped around in wording, like "Packaging_Electronics" is meant to be "Electronics_Packaging". Once again using the mutate and str_replace_all functions we were able to ensure that all were changed to only have seven categories.

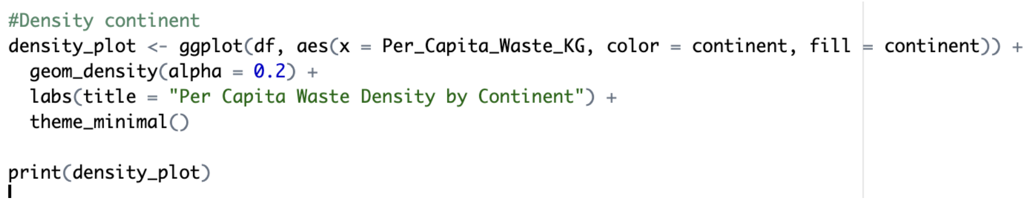
Analysis

Visualizations

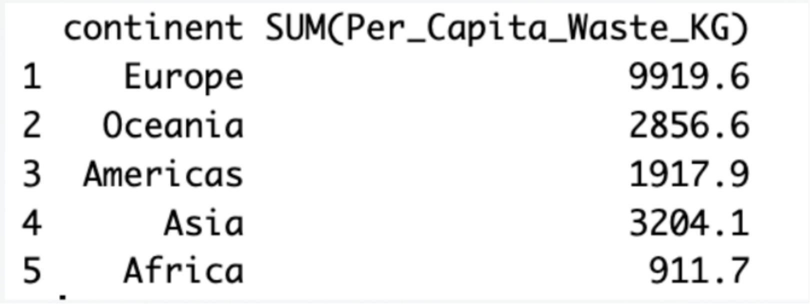
Visualization 1 : Density Chart

Using the Density chart, we decided to look into how much each continent's population was contributing to Plastic waste based on per capita. We can see that most are closer to the 0-250 range, but we can easily see there are many outliers, specifically for Europe, Oceania, Asia, and some of the Americans. We can't tell which specific countries are those outliers in this chart but we can assume that it is those major global power countries as they have more purchasing power that enables them to buy more while simultaneously contributing to more plastic waste. There are also other issues like if the country's economy is stimulated by manufacturing for example. Our question was *Which continent has the highest Per Capita Waste?*, which in this case is Europe.





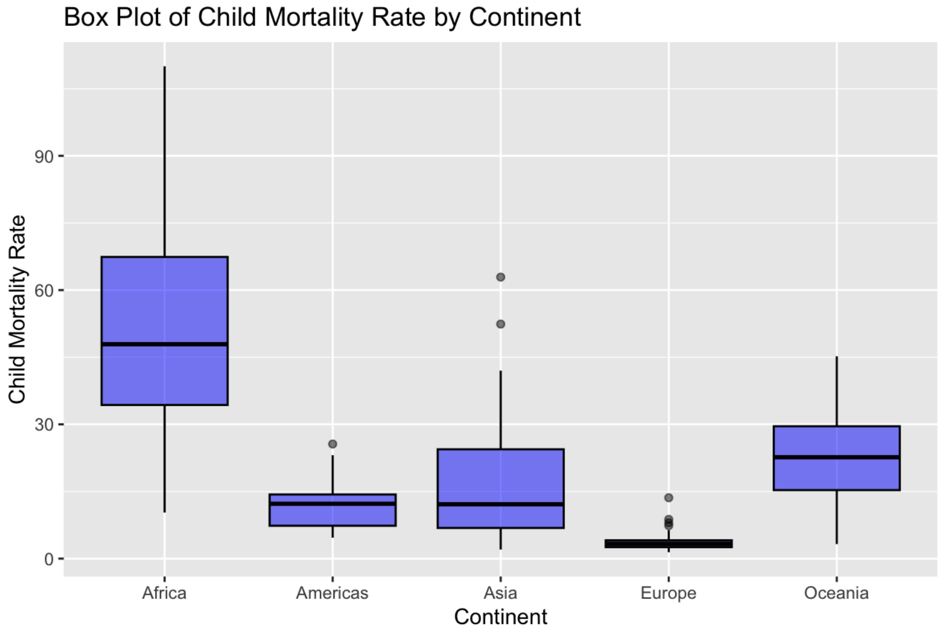
```
#Density continent
density_plot <- ggplot(df, aes(x = Per_Capita_Waste_KG, color = continent, fill = continent)) +
  geom_density(alpha = 0.2) +
  labs(title = "Per Capita Waste Density by Continent") +
  theme_minimal()

print(density_plot)
```

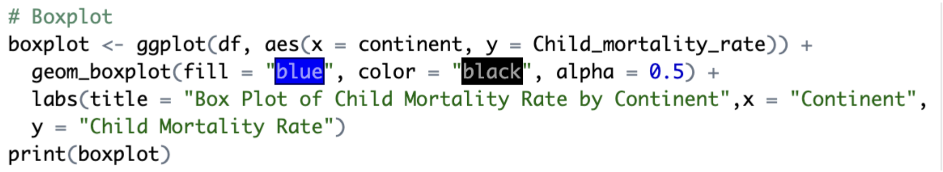


```
  continent SUM(Per_Capita_Waste_KG)
1    Europe                   9919.6
2   Oceania                   2856.6
3  Americas                   1917.9
4      Asia                   3204.1
5    Africa                    911.7
```

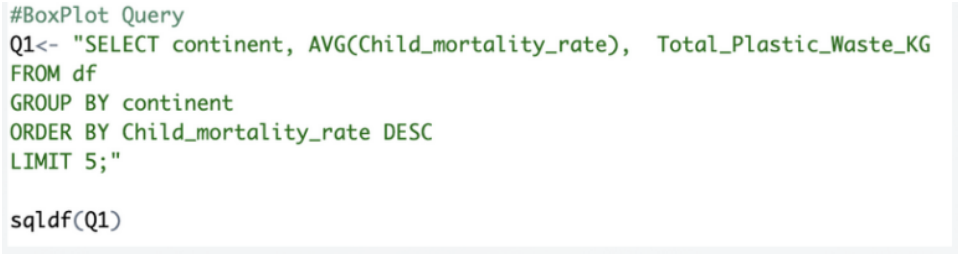
Visualization 2: Box Plot

Our box plot shows how Child Mortality Rate is shown by continent and we can see that, unfortunately, Africa has a higher median. Based on this observation we can assume that the cause of deaths can potentially lead to less economic power. Less people, means less workers and a far less stimulated economy. This also raises the question, does plastic waste lead to a higher Child Mortality rate? We will further investigate in our next query and investigation, but our question for this is: *Does more plastic waste lead to a higher child mortality rate?* Our answer was No! Africa has little plastic waste but has a higher child mortality rate. Americans have the highest plastic waste yet have the second lowest mortality rate, so this shows no correlation here.

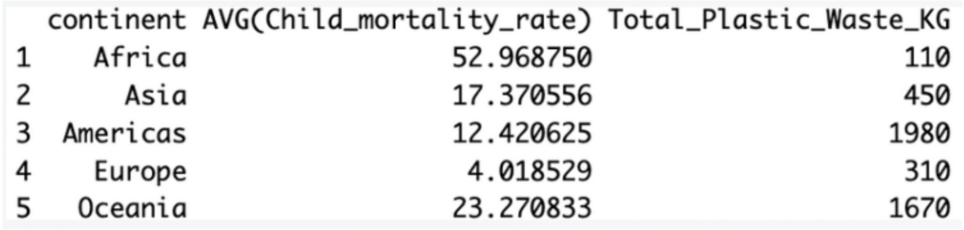




```
# Boxplot
boxplot <- ggplot(df, aes(x = continent, y = Child_mortality_rate)) +
  geom_boxplot(fill = "blue", color = "black", alpha = 0.5) +
  labs(title = "Box Plot of Child Mortality Rate by Continent",x = "Continent",
  y = "Child Mortality Rate")
print(boxplot)
```



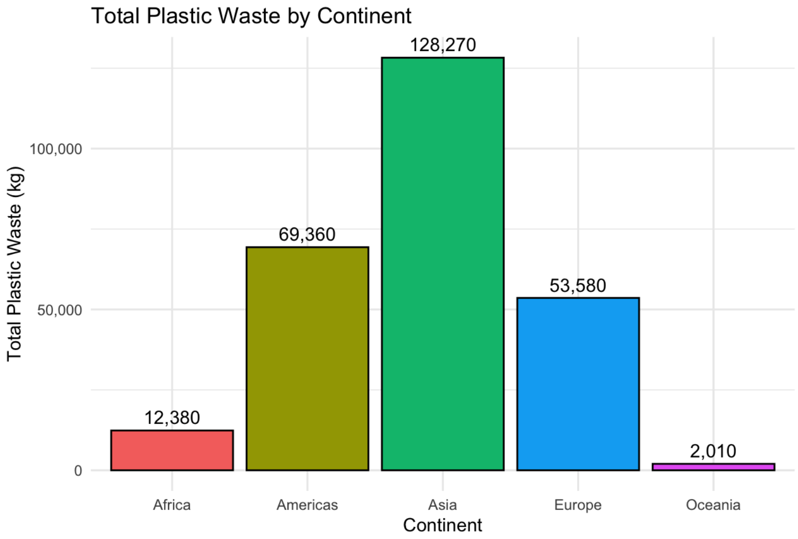
```
#BoxPlot Query
Q1<- "SELECT continent, AVG(Child_mortality_rate),  Total_Plastic_Waste_KG
FROM df
GROUP BY continent
ORDER BY Child_mortality_rate DESC
LIMIT 5;"

sqldf(Q1)
```

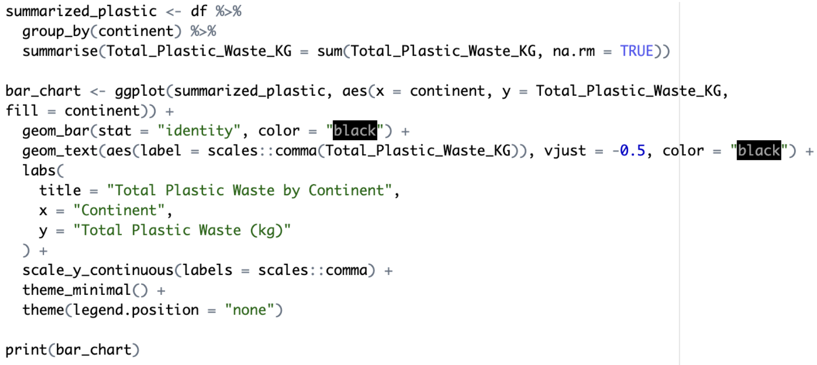


```
  continent AVG(Child_mortality_rate) Total_Plastic_Waste_KG
1    Africa                 52.968750                    110
2      Asia                 17.370556                    450
3  Americas                 12.420625                   1980
4    Europe                  4.018529                    310
5   Oceania                 23.270833                   1670
```

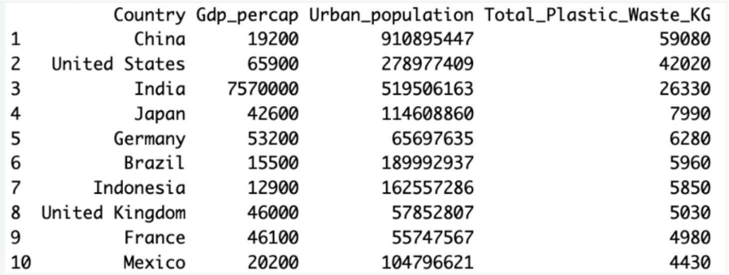
Visualization 3: Bar Chart

Our barchart shows our continents total plastic waste which we can see Asia has the highest. Referring back to our second query and visualization, we can infer that this might be due to the nature of Asia's economy which relies on manufacturing. Our question is *Does a larger population equate to more waste?* Our answer is, not necessarily, we can see that our total plastic waste isn't caused by a larger population either as India has a higher population that the United States, but India is placed third in total plastic waste. We are unsure as to how or why they are beating out the US but we know it isn't due to population size.

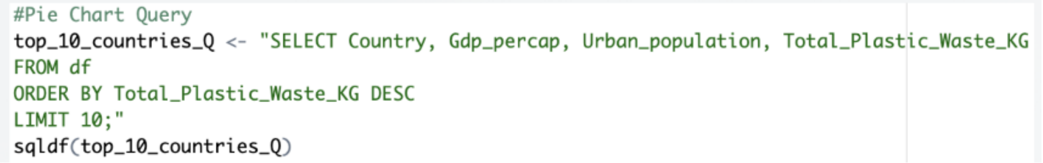




```
summarized_plastic <- df %>%
  group_by(continent) %>%
  summarise(Total_Plastic_Waste_KG = sum(Total_Plastic_Waste_KG, na.rm = TRUE))

bar_chart <- ggplot(summarized_plastic, aes(x = continent, y = Total_Plastic_Waste_KG,
fill = continent)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = scales::comma(Total_Plastic_Waste_KG)), vjust = -0.5, color = "black") +
  labs(
    title = "Total Plastic Waste by Continent",
    x = "Continent",
    y = "Total Plastic Waste (kg)"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(legend.position = "none")

print(bar_chart)
```



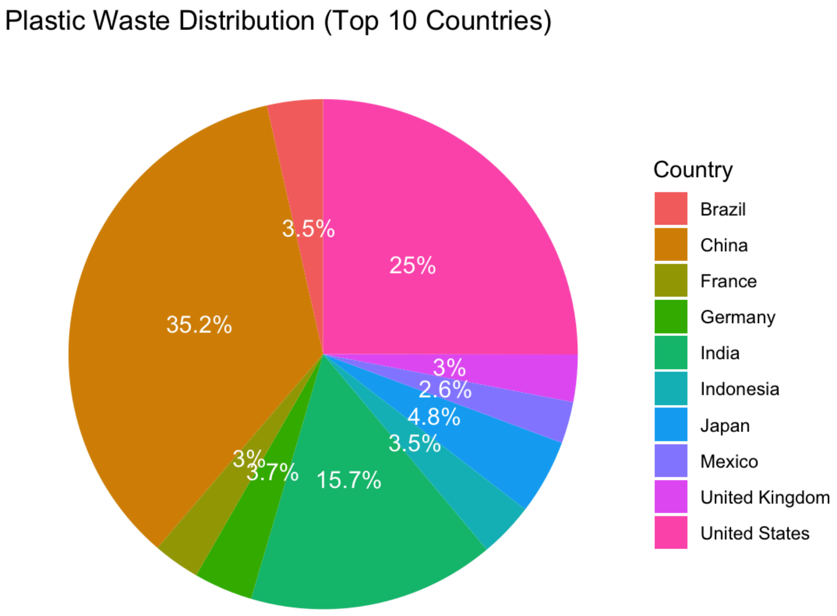
```
          Country Gdp_percap Urban_population Total_Plastic_Waste_KG
1           China      19200        910895447                  59080
2   United States      65900        278977409                  42020
3           India    7570000        519506163                  26330
4           Japan      42600        114608860                   7990
5         Germany      53200         65697635                   6280
6          Brazil      15500        189992937                   5960
7       Indonesia      12900        162557286                   5850
8  United Kingdom      46000         57852807                   5030
9          France      46100         55747567                   4980
10         Mexico      20200        104796621                   4430
```

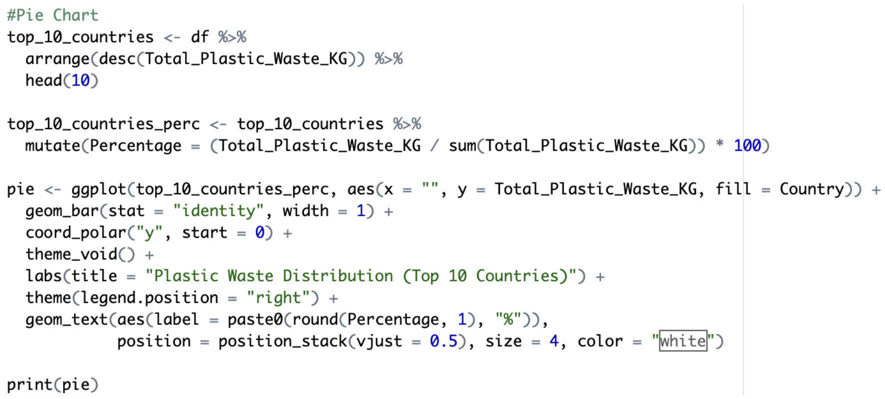


```
#Pie Chart Query
top_10_countries_Q <- "SELECT Country, Gdp_percap, Urban_population, Total_Plastic_Waste_KG
FROM df
ORDER BY Total_Plastic_Waste_KG DESC
LIMIT 10;"
sqldf(top_10_countries_Q)
```

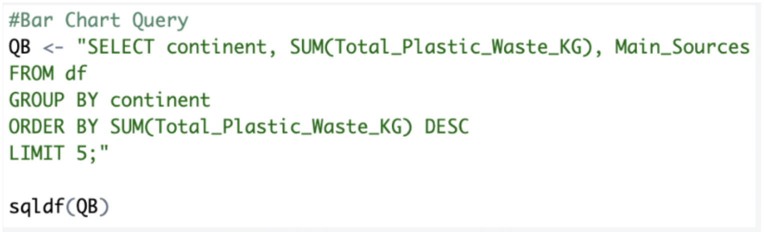
Visualization 4: Pie Chart

Our Pie Chart was created to show our top 10 countries with the most total waste. Yet again, China leads at the top which is, as we've previously stated and now seen, manufacturing and Consumer Packaging. If we look at our query below, Our question comes to: *What is the main source causing plastic waste?* Our answer leads to Consumer Packaging, and that is across the board for all continents. So this isn't a single region's problem, but instead it’s a global problem leading to lots of plastic waste.

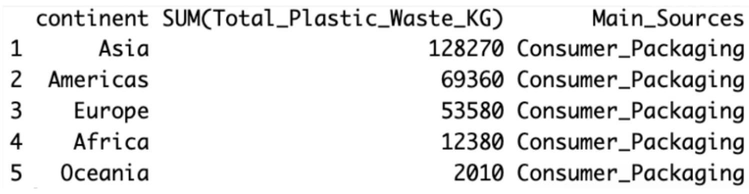




```
#Pie Chart
top_10_countries <- df %>%
  arrange(desc(Total_Plastic_Waste_KG)) %>%
  head(10)

top_10_countries_perc <- top_10_countries %>%
  mutate(Percentage = (Total_Plastic_Waste_KG / sum(Total_Plastic_Waste_KG)) * 100)

pie <- ggplot(top_10_countries_perc, aes(x = "", y = Total_Plastic_Waste_KG, fill = Country)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  theme_void() +
  labs(title = "Plastic Waste Distribution (Top 10 Countries)") +
  theme(legend.position = "right") +
  geom_text(aes(label = paste0(round(Percentage, 1), "%")),
            position = position_stack(vjust = 0.5), size = 4, color = "white")

print(pie)
```



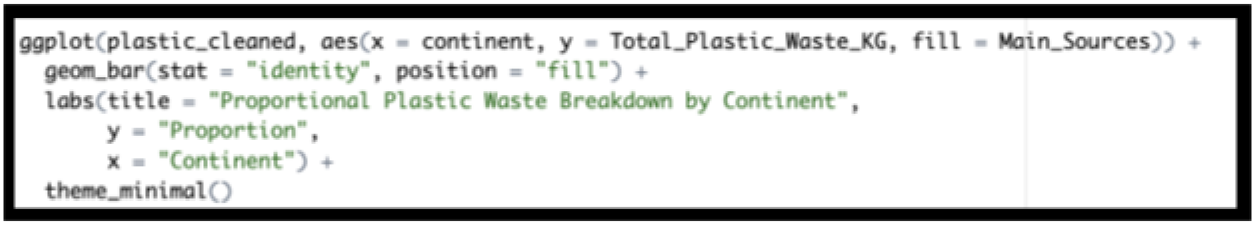
```
#Bar Chart Query
QB <- "SELECT continent, SUM(Total_Plastic_Waste_KG), Main_Sources
FROM df
GROUP BY continent
ORDER BY SUM(Total_Plastic_Waste_KG) DESC
LIMIT 5;"

sqldf(QB)
```



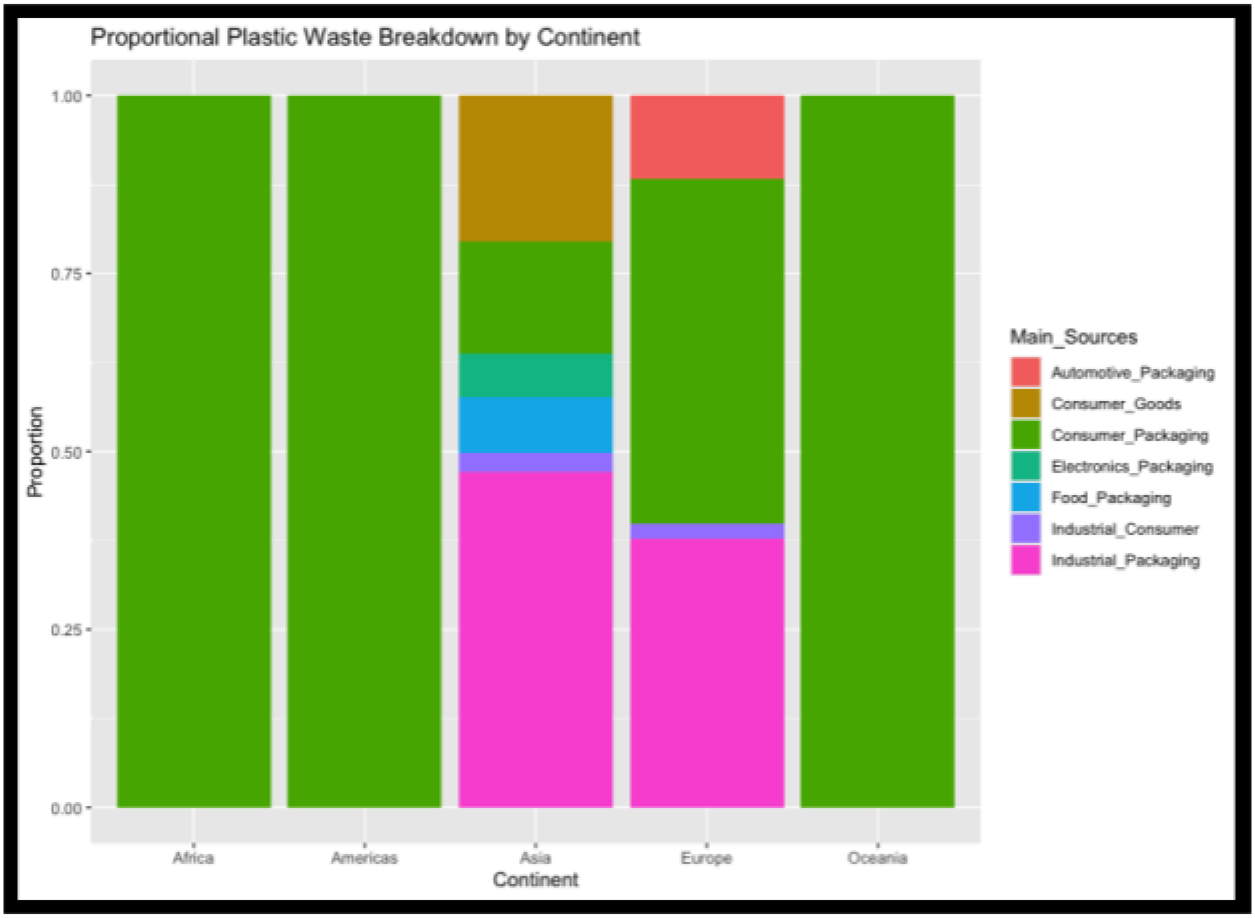
```
  continent SUM(Total_Plastic_Waste_KG)        Main_Sources
1      Asia                      128270 Consumer_Packaging
2  Americas                       69360 Consumer_Packaging
3    Europe                       53580 Consumer_Packaging
4    Africa                       12380 Consumer_Packaging
5   Oceania                        2010 Consumer_Packaging
```

Visualization 5: Stacked Bar Graph

To understand the breakdown of plastic waste sources of every region we created the following stacked bar graph. Each color represents the source of plastic and the proportion is of the total plastic waste grouped by the country data that we had from each continent. As displayed in the visualization the Americas, Africa, and Oceania have 100% of total plastic waste generated from consumer packaging. With Asia and Europe having more diversity in sources of plastic waste, with Asian countries having more industrial packaging being the main source of plastic at almost 50%. This helps visualize the responsibility that each region has in generating plastic waste because this could be tied to the leading industries in these regions or the imported goods.

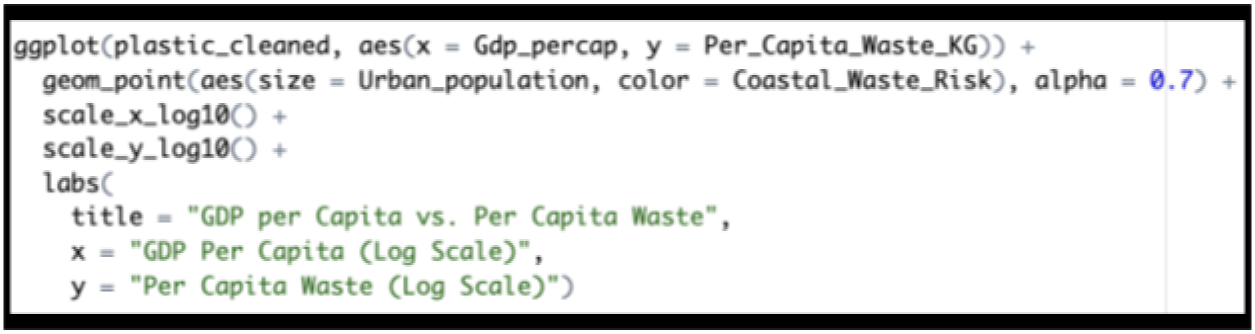


```
ggplot(plastic_cleaned, aes(x = continent, y = Total_Plastic_Waste_KG, fill = Main_Sources)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Proportional Plastic Waste Breakdown by Continent",
       y = "Proportion",
       x = "Continent") +
  theme_minimal()
```

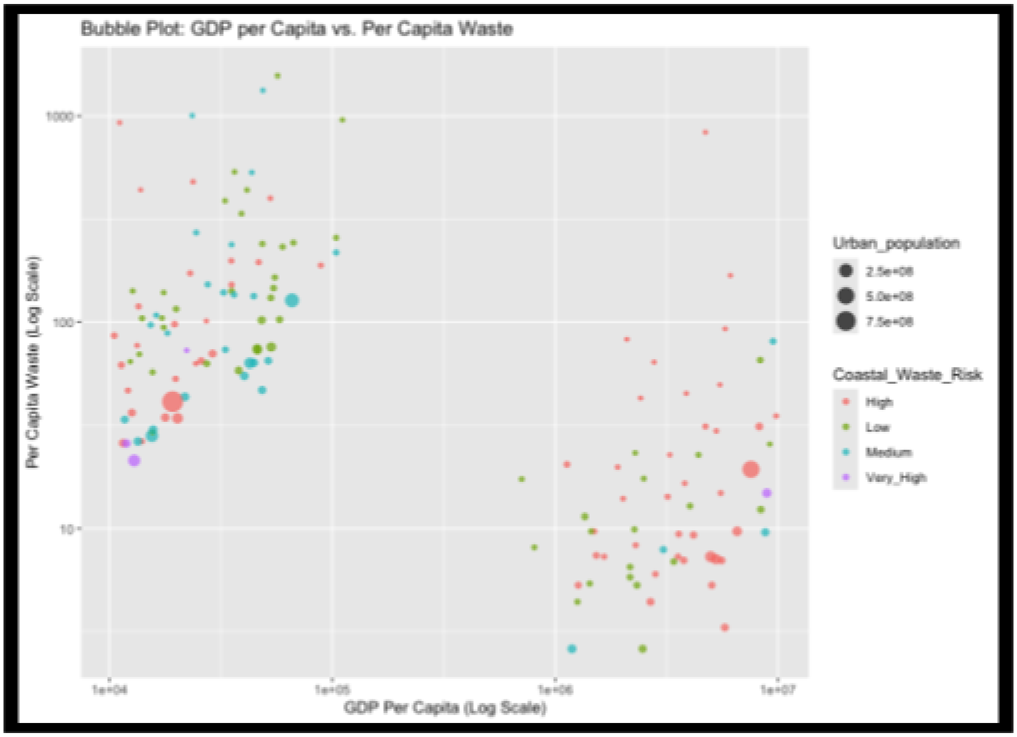


<u>Visualization 6: Bubble Plot</u>

This visualization was created to understand what kind of relationship there was between GDP per capita and per capita waste production, with each data point representing a different country. Based on this there is a clear split between the 2, but there is a negative possible relationship between the two variables. This would make sense as we would expect there to be a higher waste production for countries with a lower GDP per capita because it represents a less economically powerful country using more plastic. The clear split in the data could be due to the fact that there were deletions of countries that we weren't able to attain data for which left us with countries with more extreme GDP per capita. The size of the points correlate with the urban population size, and the color is correlated with the coastal waste risk level assigned to the country. The distribution of urban population and coastal waste risk don't seem as heavily correlated with the other two variables, but that might be because it has more to do with geographical location than with the two variables we have chosen in this visualization.

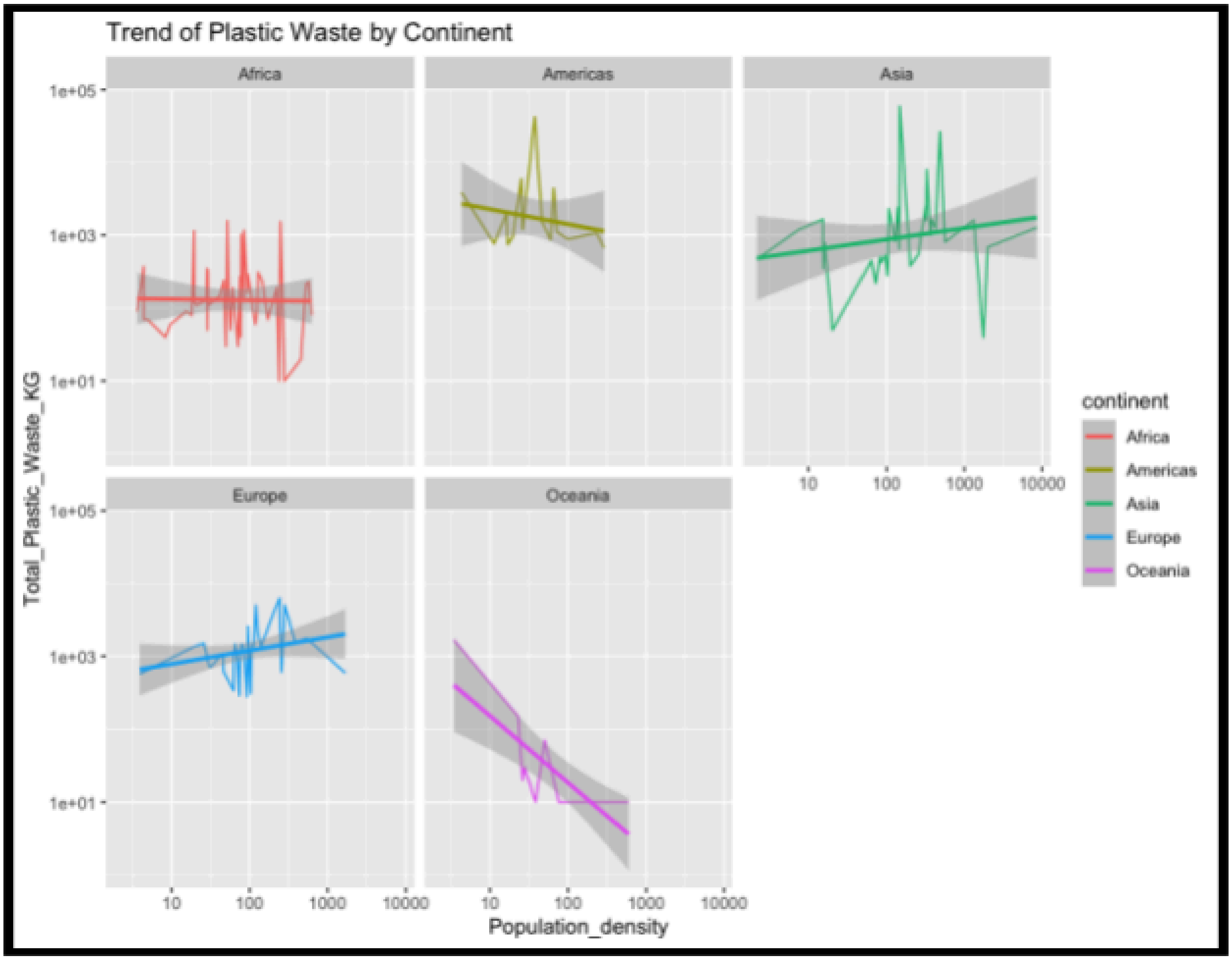


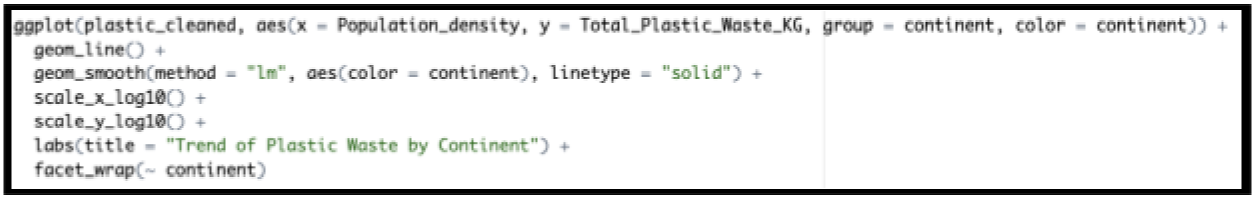
```
ggplot(plastic_cleaned, aes(x = Gdp_percap, y = Per_Capita_Waste_KG)) +
  geom_point(aes(size = Urban_population, color = Coastal_Waste_Risk), alpha = 0.7) +
  scale_x_log10() +
  scale_y_log10() +
  labs(
    title = "GDP per Capita vs. Per Capita Waste",
    x = "GDP Per Capita (Log Scale)",
    y = "Per Capita Waste (Log Scale)")
```



Visualization 7: Line Graph

The use of a line graph and regression line was to demonstrate any relationship between the total amount of plastic waste and population density, categorized by continent. The regression line was used to point out a more obvious trend amongst the data since there was so much variation between countries within each continent. The graph was also scaled with logarithmic axes to help identify clearer patterns across varying scales of both variables. As the visualization demonstrates that the regression line fitted in each graph is different, not indicating a global trend. This data again can be skewed by how many countries from each continent we were able to keep within our dataset during the cleaning process.

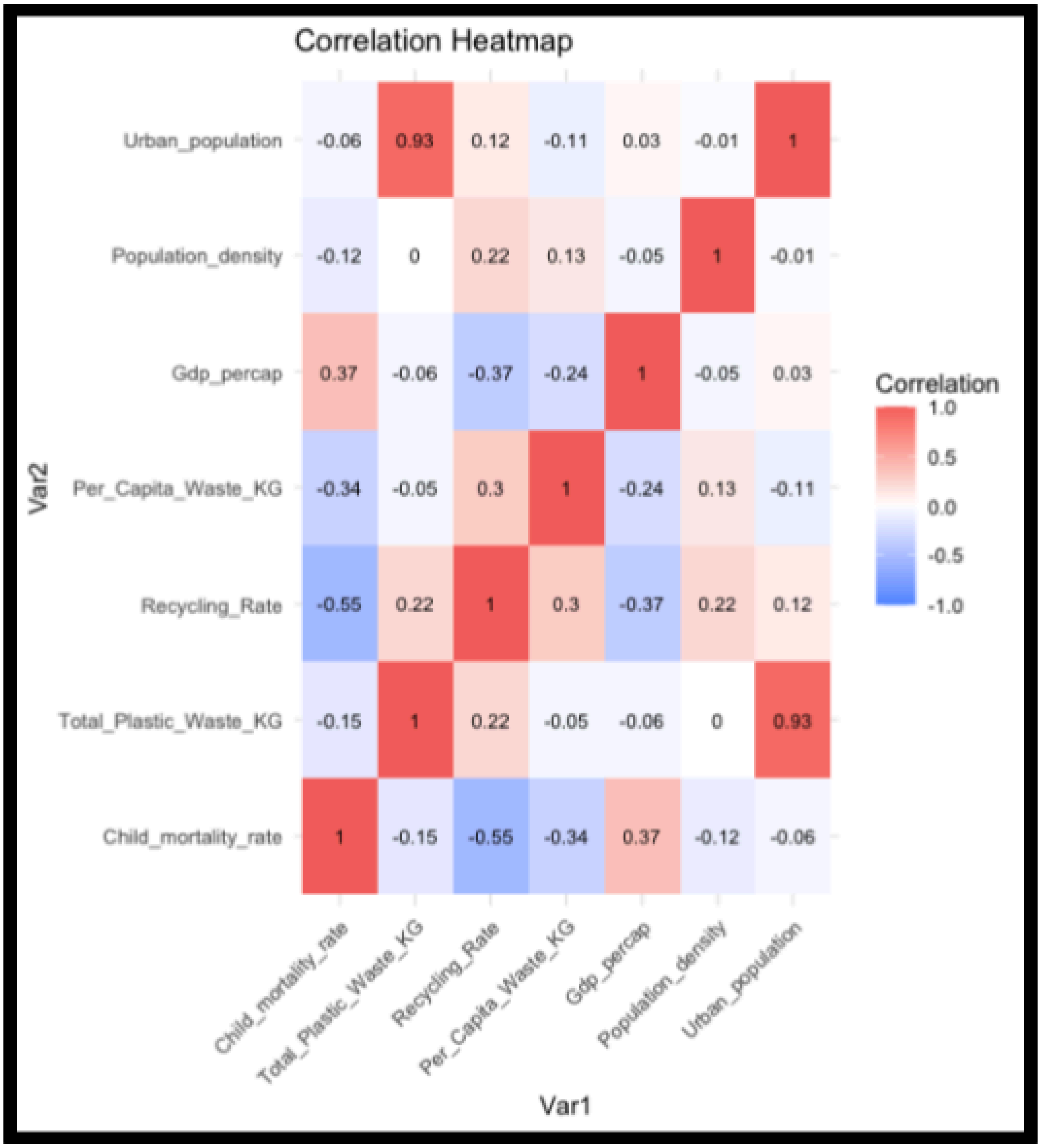


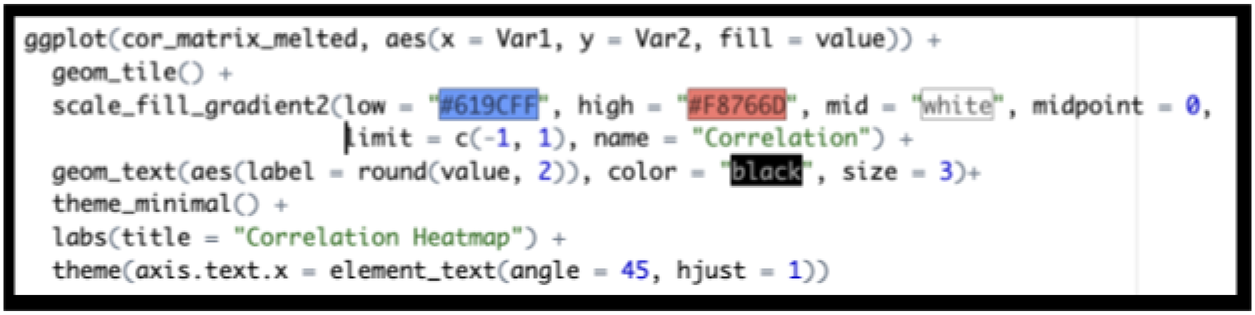


```
ggplot(plastic_cleaned, aes(x = Population_density, y = Total_Plastic_Waste_KG, group = continent, color = continent)) +
  geom_line() +
  geom_smooth(method = "lm", aes(color = continent), linetype = "solid") +
  scale_x_log10() +
  scale_y_log10() +
  labs(title = "Trend of Plastic Waste by Continent") +
  facet_wrap(~ continent)
```

<u>Visualization 8: Heatmap</u>

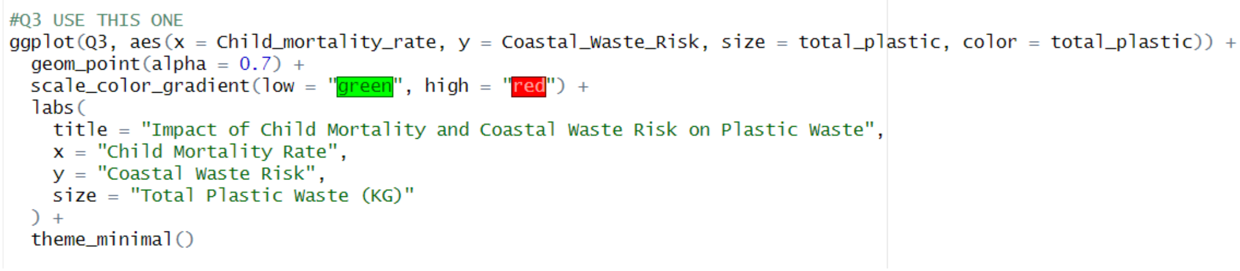
This following visualization is a correlation heatmap that displays the correlation between all numeric values in our dataset. With the values in red being closer to 1 meaning that there is a higher correlation between the two variables. In this heat map we can see that urban population and total plastic waste has a high positive correlation meaning that there increases in either is associated with increase in the other. This is also expected because urbanization means that there are establishments, like restaurants, that contribute more to plastic waste. Recycling rate and child mortality rate seem to demonstrate a slightly high negative correlation which could mean that a higher recycling rate can be associated with a lower child mortality rate.



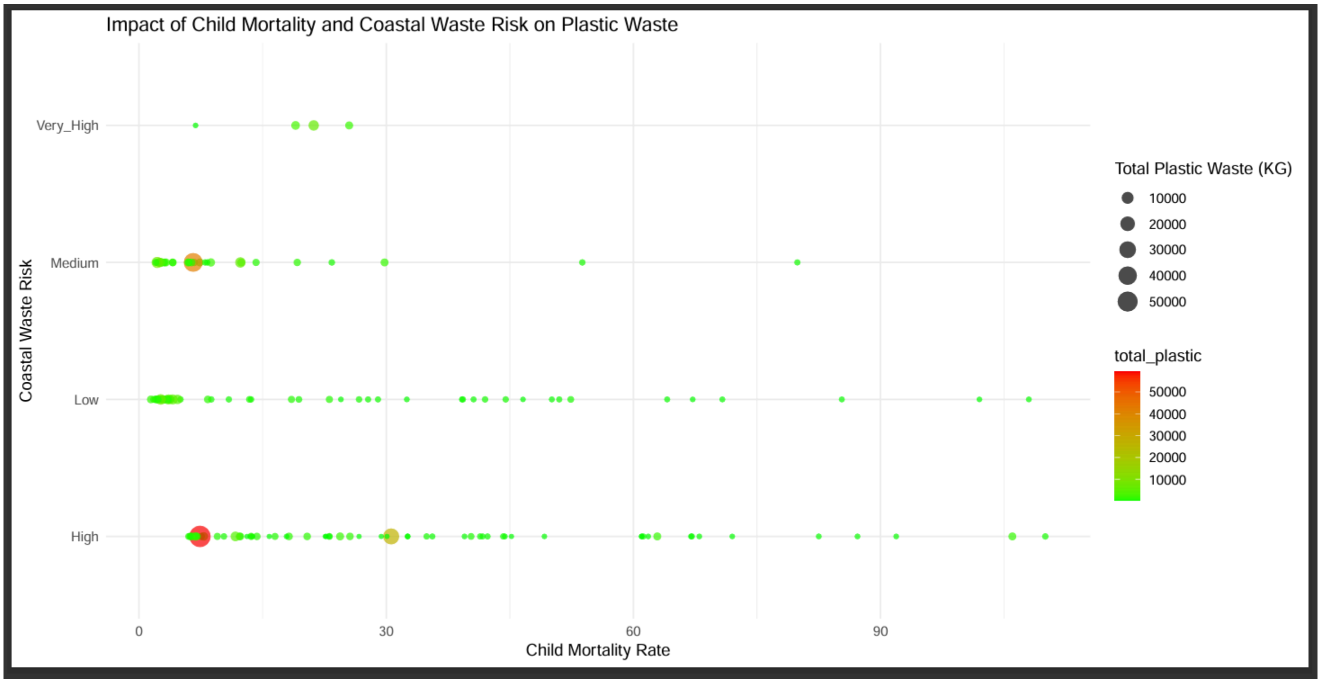


```
ggplot(cor_matrix_melted, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "#619CFF", high = "#F8766D", mid = "white", midpoint = 0,
                       limit = c(-1, 1), name = "Correlation") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 3)+
  theme_minimal() +
  labs(title = "Correlation Heatmap") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

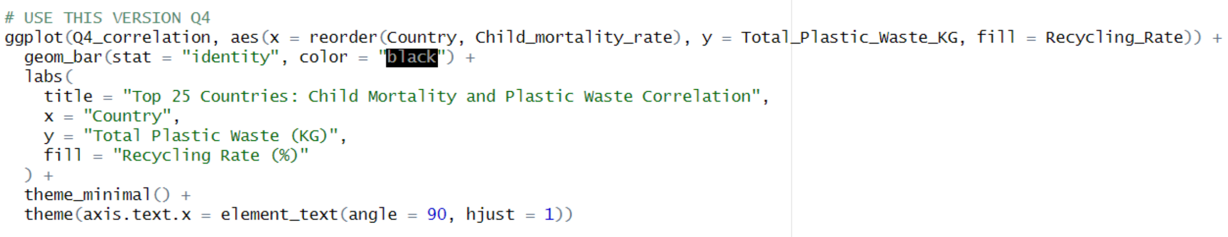
Visualization 9: Bubble Chart



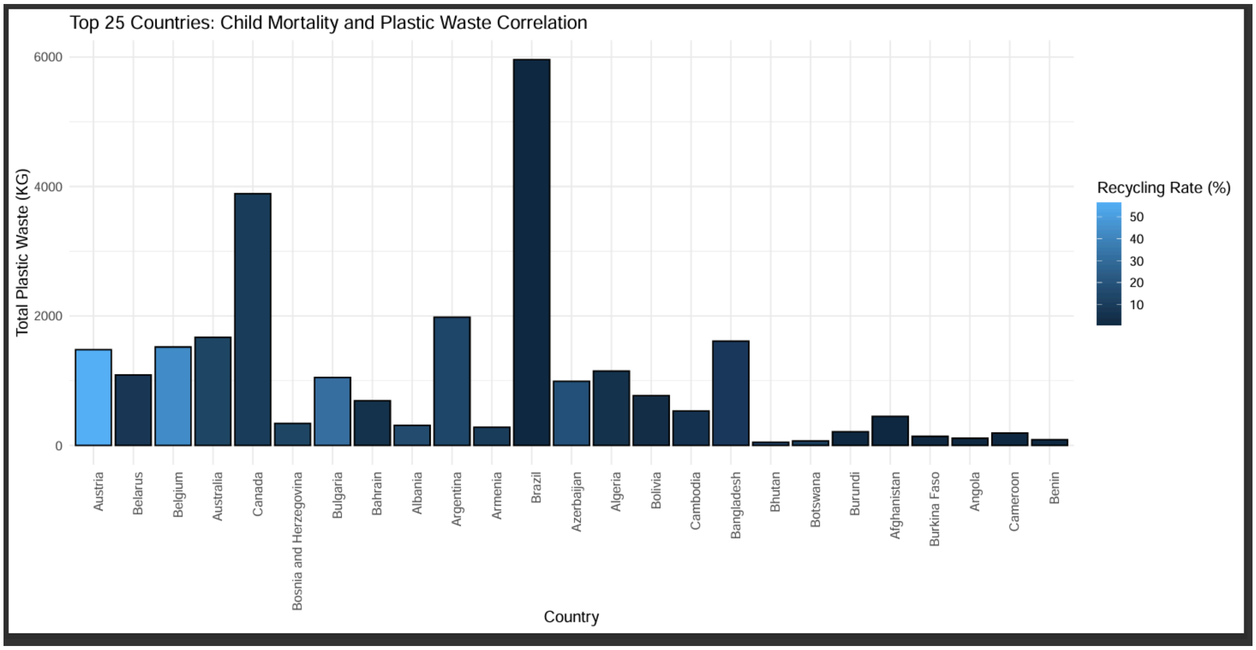
```
#Q3 USE THIS ONE
ggplot(Q3, aes(x = Child_mortality_rate, y = Coastal_Waste_Risk, size = total_plastic, color = total_plastic)) +
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "green", high = "red") +
  labs(
    title = "Impact of Child Mortality and Coastal Waste Risk on Plastic Waste",
    x = "Child Mortality Rate",
    y = "Coastal Waste Risk",
    size = "Total Plastic Waste (KG)"
  ) +
  theme_minimal()
```



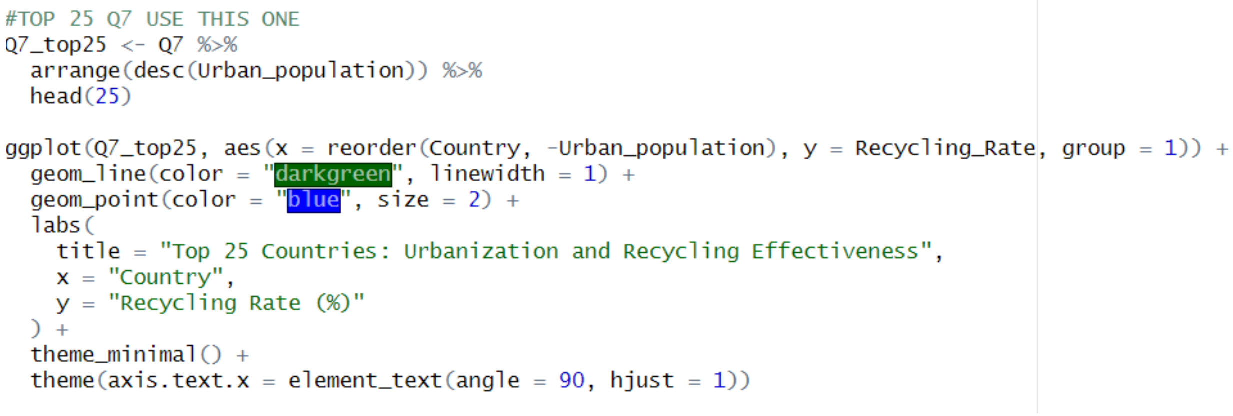
Visualization 10: Bar Chart



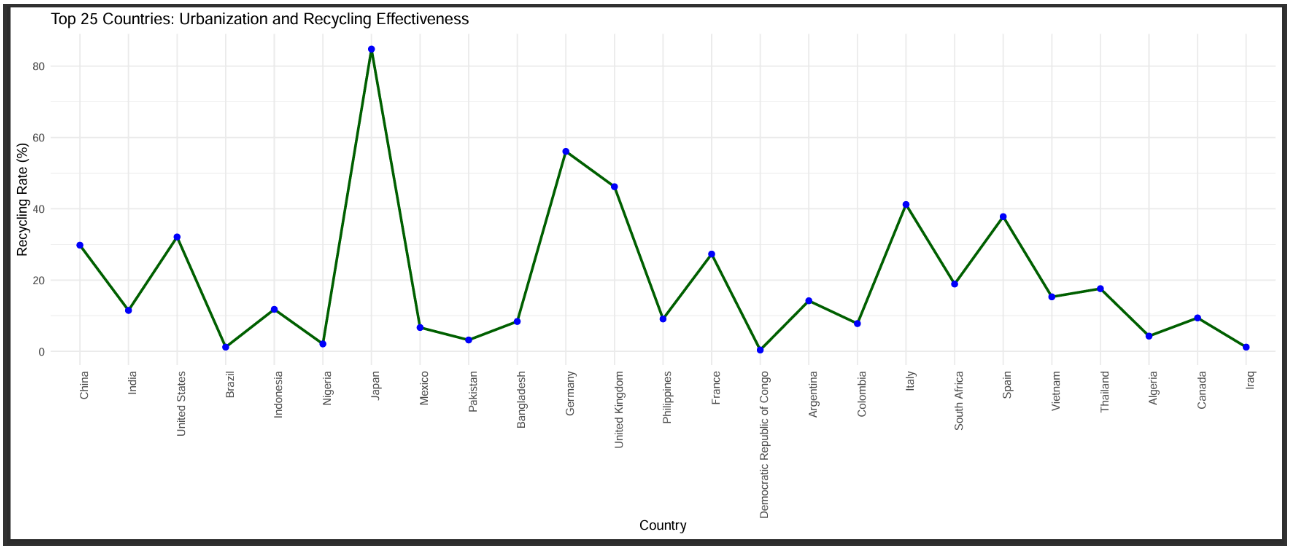
```
# USE THIS VERSION Q4
ggplot(Q4_correlation, aes(x = reorder(Country, Child_mortality_rate), y = Total_Plastic_Waste_KG, fill = Recycling_Rate)) +
  geom_bar(stat = "identity", color = "black") +
  labs(
    title = "Top 25 Countries: Child Mortality and Plastic Waste Correlation",
    x = "Country",
    y = "Total Plastic Waste (KG)",
    fill = "Recycling Rate (%)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



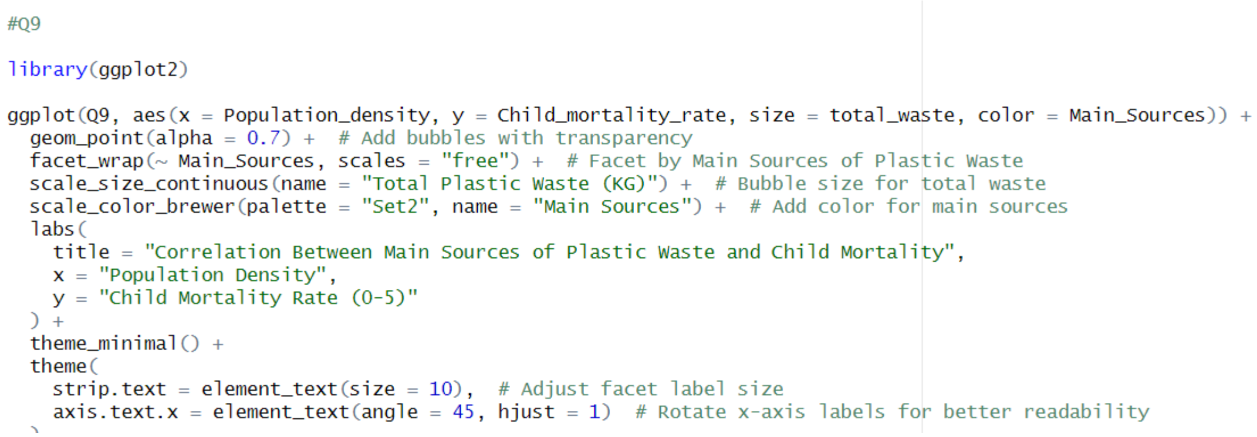
Visualization 11: Line Chart



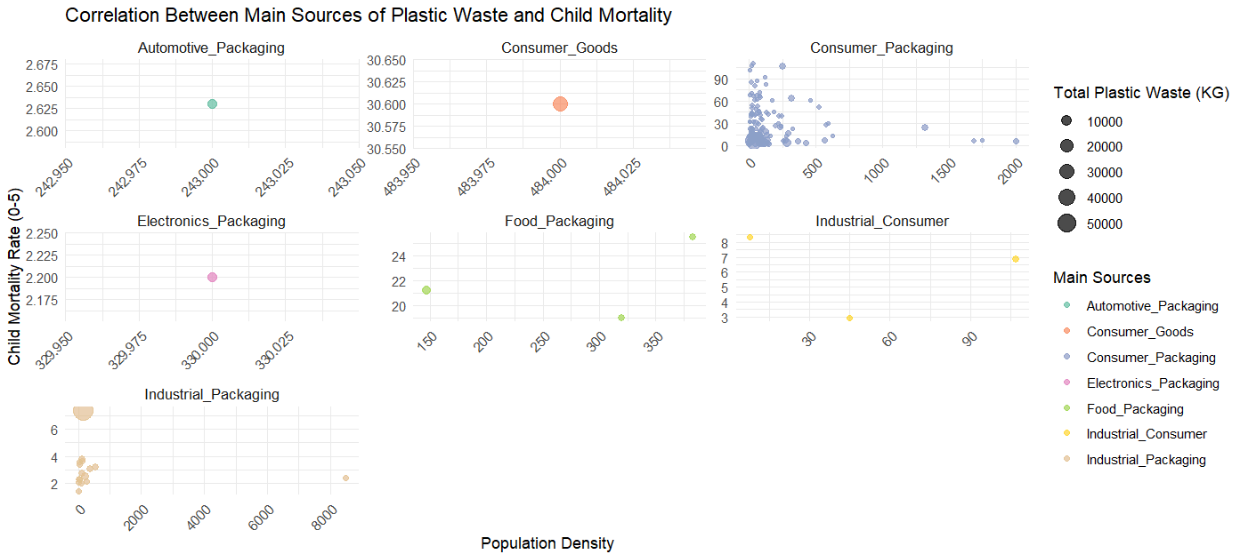
```
#TOP 25 Q7 USE THIS ONE
Q7_top25 <- Q7 %>%
  arrange(desc(Urban_population)) %>%
  head(25)

ggplot(Q7_top25, aes(x = reorder(Country, -Urban_population), y = Recycling_Rate, group = 1)) +
  geom_line(color = "darkgreen", linewidth = 1) +
  geom_point(color = "blue", size = 2) +
  labs(
    title = "Top 25 Countries: Urbanization and Recycling Effectiveness",
    x = "Country",
    y = "Recycling Rate (%)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Visualization 12: Facet Bubble Plot

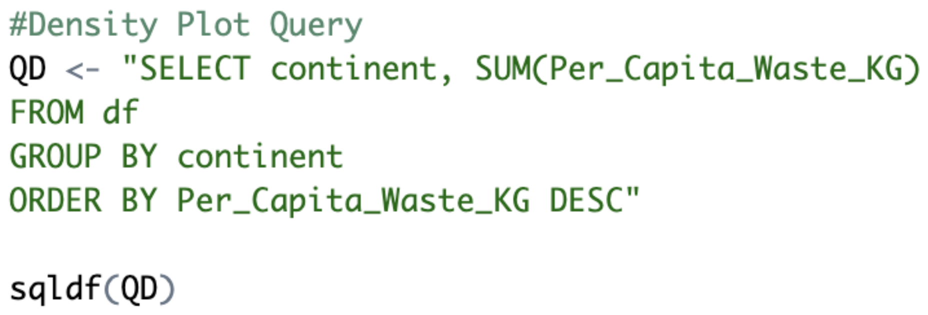


```
#Q9

library(ggplot2)

ggplot(Q9, aes(x = Population_density, y = Child_mortality_rate, size = total_waste, color = Main_Sources)) +
  geom_point(alpha = 0.7) +  # Add bubbles with transparency
  facet_wrap(~ Main_Sources, scales = "free") +  # Facet by Main Sources of Plastic Waste
  scale_size_continuous(name = "Total Plastic Waste (KG)") +  # Bubble size for total waste
  scale_color_brewer(palette = "Set2", name = "Main Sources") +  # Add color for main sources
  labs(
    title = "Correlation Between Main Sources of Plastic Waste and Child Mortality",
    x = "Population Density",
    y = "Child Mortality Rate (0-5)"
  ) +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 10),  # Adjust facet label size
    axis.text.x = element_text(angle = 45, hjust = 1)  # Rotate x-axis labels for better readability
```

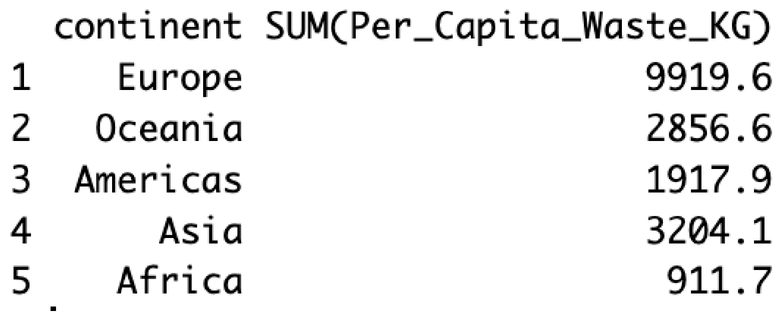


Queries

Query 1

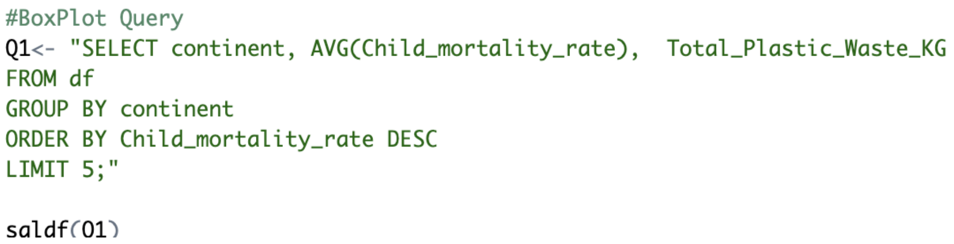


```
#Density Plot Query
QD <- "SELECT continent, SUM(Per_Capita_Waste_KG)
FROM df
GROUP BY continent
ORDER BY Per_Capita_Waste_KG DESC"

sqldf(QD)
```

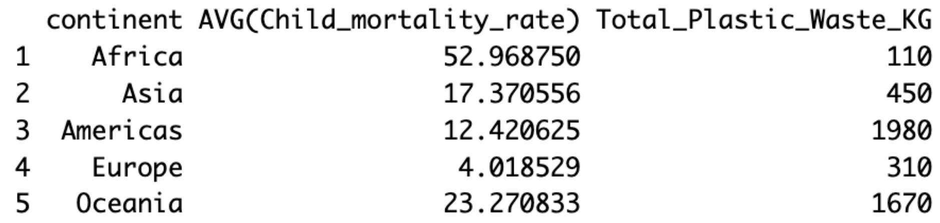


```
  continent SUM(Per_Capita_Waste_KG)
1    Europe                   9919.6
2   Oceania                   2856.6
3  Americas                   1917.9
4      Asia                   3204.1
5    Africa                    911.7
```

Query 2

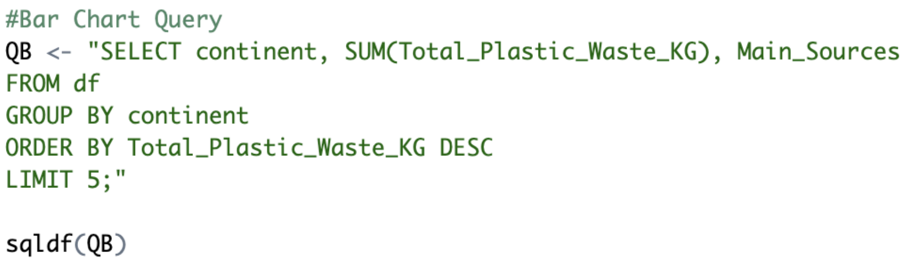


```
#BoxPlot Query
Q1<- "SELECT continent, AVG(Child_mortality_rate),  Total_Plastic_Waste_KG
FROM df
GROUP BY continent
ORDER BY Child_mortality_rate DESC
LIMIT 5;"

sqldf(Q1)
```

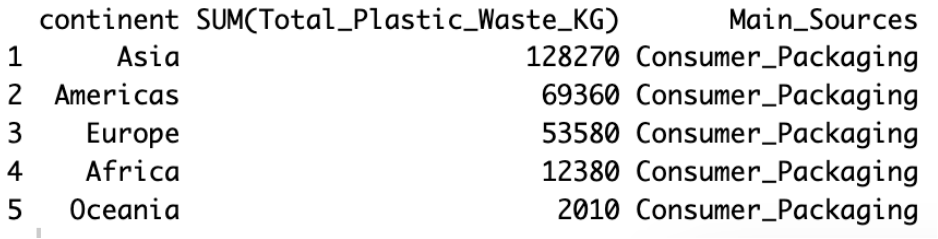


```
  continent AVG(Child_mortality_rate) Total_Plastic_Waste_KG
1    Africa                 52.968750                    110
2      Asia                 17.370556                    450
3  Americas                 12.420625                   1980
4    Europe                  4.018529                    310
5   Oceania                 23.270833                   1670
```

Query 3

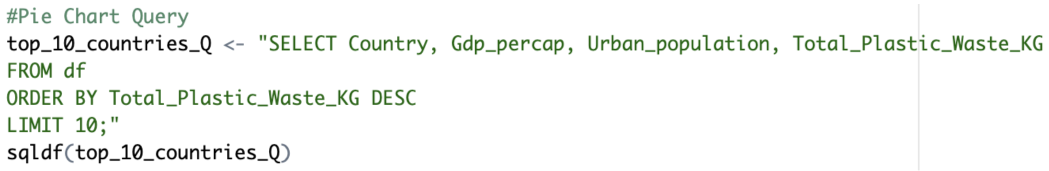


```
#Bar Chart Query
QB <- "SELECT continent, SUM(Total_Plastic_Waste_KG), Main_Sources
FROM df
GROUP BY continent
ORDER BY Total_Plastic_Waste_KG DESC
LIMIT 5;"

sqldf(QB)
```

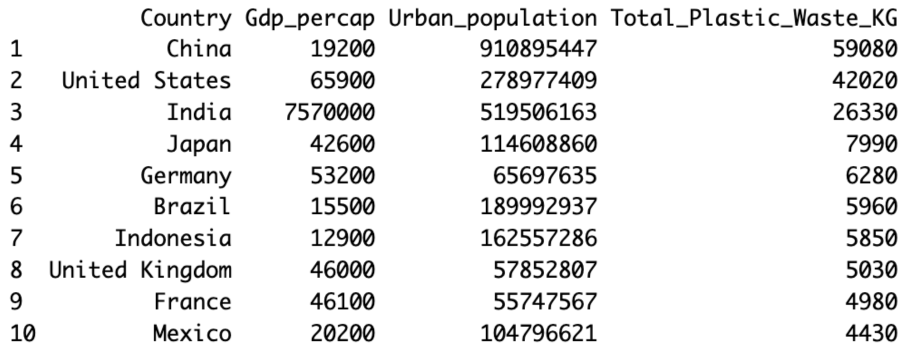


```
  continent SUM(Total_Plastic_Waste_KG)       Main_Sources
1      Asia                      128270 Consumer_Packaging
2  Americas                       69360 Consumer_Packaging
3    Europe                       53580 Consumer_Packaging
4    Africa                       12380 Consumer_Packaging
5   Oceania                        2010 Consumer_Packaging
```

Query 4



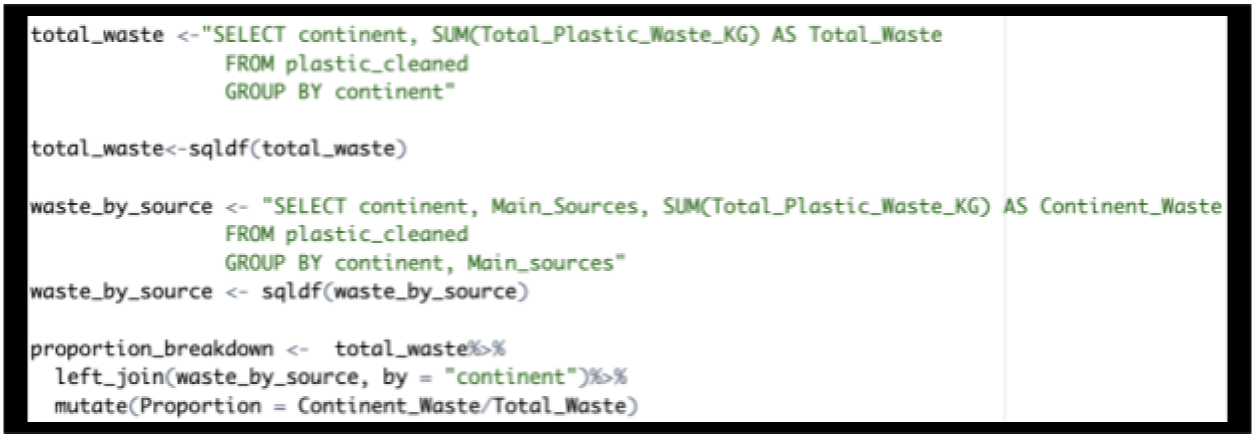
```
#Pie Chart Query
top_10_countries_Q <- "SELECT Country, Gdp_percap, Urban_population, Total_Plastic_Waste_KG
FROM df
ORDER BY Total_Plastic_Waste_KG DESC
LIMIT 10;"
sqldf(top_10_countries_Q)
```



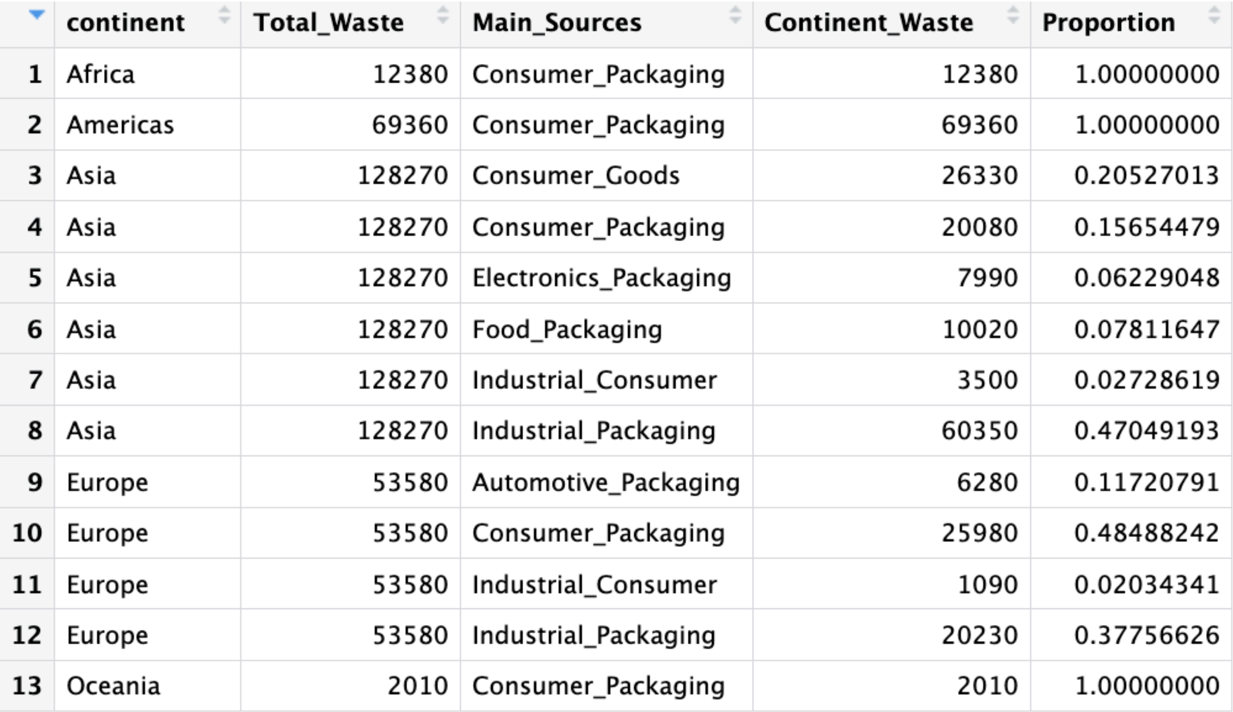
```
          Country Gdp_percap Urban_population Total_Plastic_Waste_KG
1           China      19200        910895447                  59080
2   United States      65900        278977409                  42020
3           India    7570000        519506163                  26330
4           Japan      42600        114608860                   7990
5         Germany      53200         65697635                   6280
6          Brazil      15500        189992937                   5960
7       Indonesia      12900        162557286                   5850
8  United Kingdom      46000         57852807                   5030
9          France      46100         55747567                   4980
10         Mexico      20200        104796621                   4430
```

Query 5: What is the proportional breakdown of plastic waste sources of each continent?

Query 5 was exploring if there was any industry or source that was prominent in contributing to plastic waste for each continental region. This was done by adding the total plastic waste and grouping it by continent and then by continent and source to be able to calculate the proportion of plastic waste sources to total waste. This helped to highlight what industries contributed most to plastic waste. This data could be used to then enforce more regulation on these industries within these regions to help mitigate the plastic waste production.



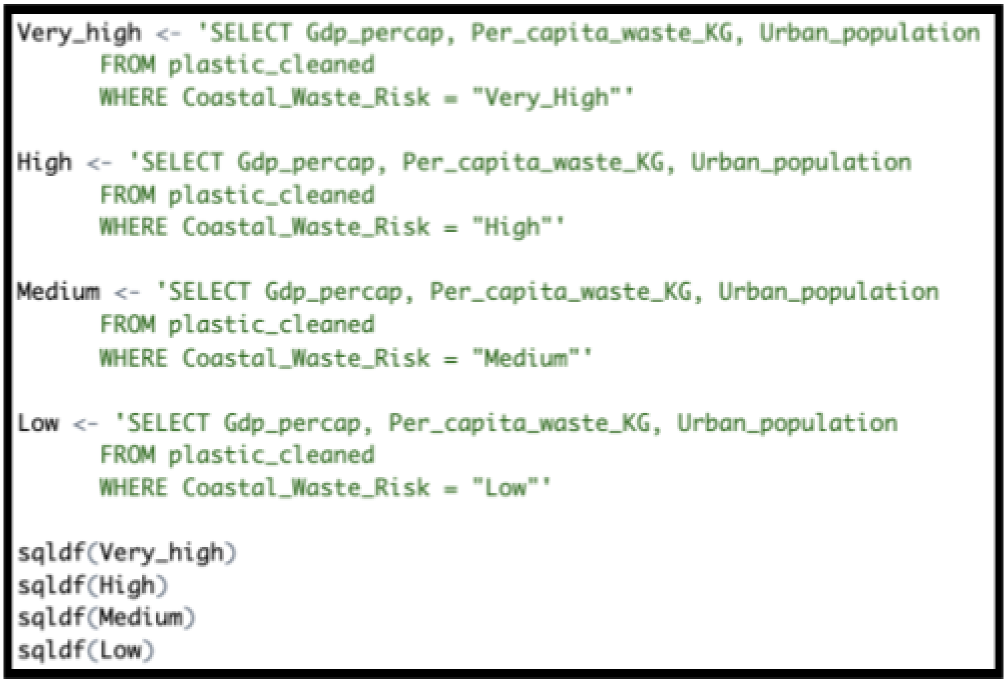
```
total_waste <-"SELECT continent, SUM(Total_Plastic_Waste_KG) AS Total_Waste
                FROM plastic_cleaned
                GROUP BY continent"

total_waste<-sqldf(total_waste)

waste_by_source <- "SELECT continent, Main_Sources, SUM(Total_Plastic_Waste_KG) AS Continent_Waste
                FROM plastic_cleaned
                GROUP BY continent, Main_sources"
waste_by_source <- sqldf(waste_by_source)

proportion_breakdown <-  total_waste%>%
  left_join(waste_by_source, by = "continent")%>%
  mutate(Proportion = Continent_Waste/Total_Waste)
```



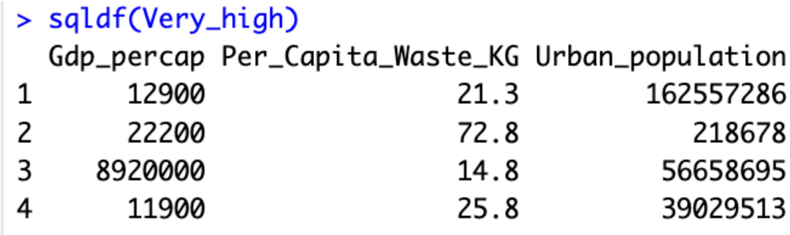
	continent	Total_Waste	Main_Sources	Continent_Waste	Proportion
1	Africa	12380	Consumer_Packaging	12380	1.00000000
2	Americas	69360	Consumer_Packaging	69360	1.00000000
3	Asia	128270	Consumer_Goods	26330	0.20527013
4	Asia	128270	Consumer_Packaging	20080	0.15654479
5	Asia	128270	Electronics_Packaging	7990	0.06229048
6	Asia	128270	Food_Packaging	10020	0.07811647
7	Asia	128270	Industrial_Consumer	3500	0.02728619
8	Asia	128270	Industrial_Packaging	60350	0.47049193
9	Europe	53580	Automotive_Packaging	6280	0.11720791
10	Europe	53580	Consumer_Packaging	25980	0.48488242
11	Europe	53580	Industrial_Consumer	1090	0.02034341
12	Europe	53580	Industrial_Packaging	20230	0.37756626
13	Oceania	2010	Consumer_Packaging	2010	1.00000000

Query 6: Is there any relationship between GDP per capita and per capita waste production, when grouped by coastal waste risk categories?

This query created multiplied tables to understand how many countries there were within these different categories of coastal waste risk and then to compare the GDP per capita and per capita waste production. There wasn't much of a connection demonstrated between the categories and the 2 other variables chosen, along with urban population size. This could be because the coastal waste risk has more to do with the geographical location of the country and the production of waste rather than the GDP. However, the hypothesis going into this is that higher GDP and lower per capita waste production would mean a low coastal waste risk category. However as seen in the corresponding visualization and in the tables below there is no real relationship between the two meaning that there could be another variable that would have more of an effect on the coastal waste risk, like geographical characterisitics.

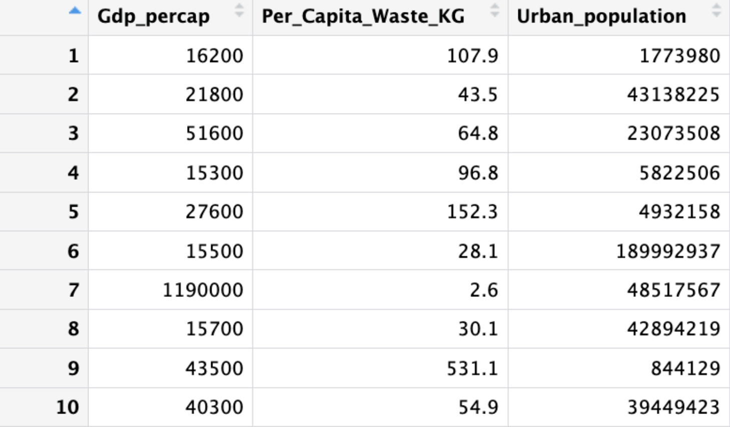


```
Very_high <- 'SELECT Gdp_percap, Per_capita_waste_KG, Urban_population
      FROM plastic_cleaned
      WHERE Coastal_Waste_Risk = "Very_High"'

High <- 'SELECT Gdp_percap, Per_capita_waste_KG, Urban_population
      FROM plastic_cleaned
      WHERE Coastal_Waste_Risk = "High"'

Medium <- 'SELECT Gdp_percap, Per_capita_waste_KG, Urban_population
      FROM plastic_cleaned
      WHERE Coastal_Waste_Risk = "Medium"'

Low <- 'SELECT Gdp_percap, Per_capita_waste_KG, Urban_population
      FROM plastic_cleaned
      WHERE Coastal_Waste_Risk = "Low"'

sqldf(Very_high)
sqldf(High)
sqldf(Medium)
sqldf(Low)
```



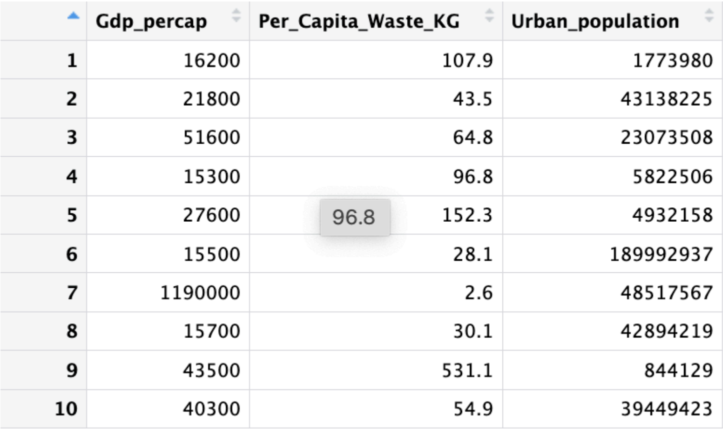
```
> sqldf(Very_high)
  Gdp_percap Per_Capita_Waste_KG Urban_population
1      12900                21.3        162557286
2      22200                72.8           218678
3    8920000                14.8         56658695
4      11900                25.8         39029513
```

High coastal waste risk



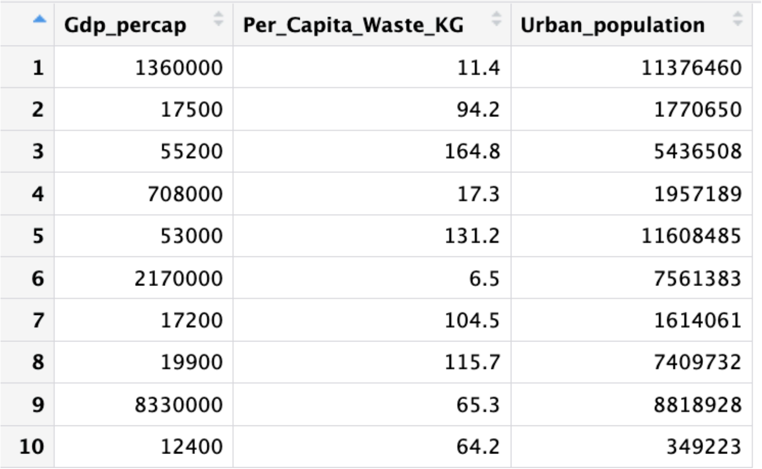
	Gdp_percap	Per_Capita_Waste_KG	Urban_population
1	16200	107.9	1773980
2	21800	43.5	43138225
3	51600	64.8	23073508
4	15300	96.8	5822506
5	27600	152.3	4932158
6	15500	28.1	189992937
7	1190000	2.6	48517567
8	15700	30.1	42894219
9	43500	531.1	844129
10	40300	54.9	39449423

Medium Coastal Waste Risk



	Gdp_percap	Per_Capita_Waste_KG	Urban_population
1	16200	107.9	1773980
2	21800	43.5	43138225
3	51600	64.8	23073508
4	15300	96.8	5822506
5	27600	152.3	4932158
6	15500	28.1	189992937
7	1190000	2.6	48517567
8	15700	30.1	42894219
9	43500	531.1	844129
10	40300	54.9	39449423

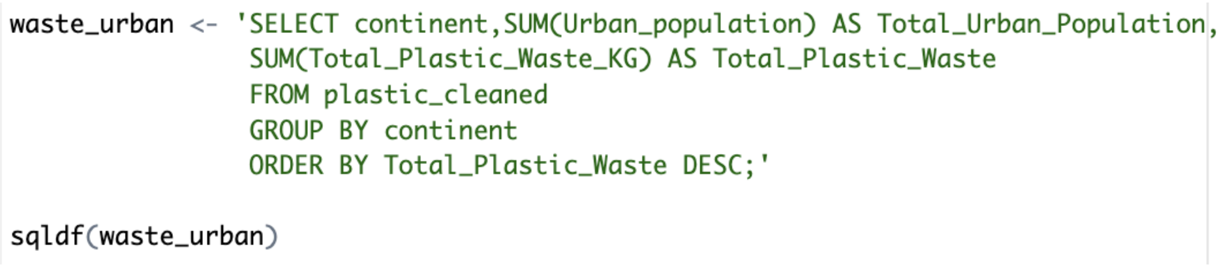
Low Coastal Waste Risk



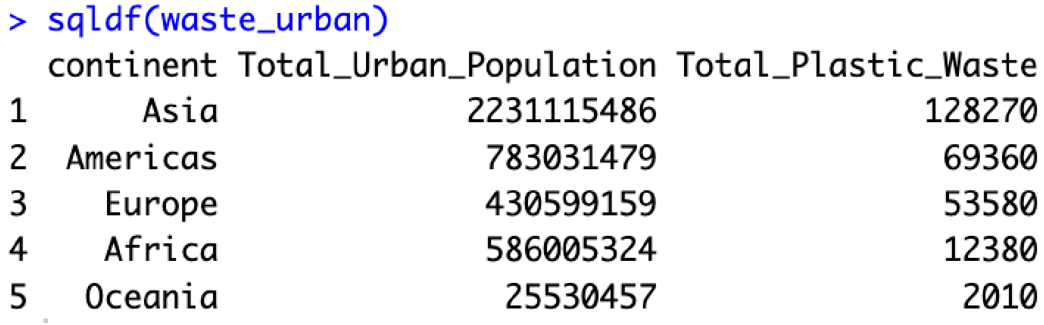
	Gdp_percap	Per_Capita_Waste_KG	Urban_population
1	1360000	11.4	11376460
2	17500	94.2	1770650
3	55200	164.8	5436508
4	708000	17.3	1957189
5	53000	131.2	11608485
6	2170000	6.5	7561383
7	17200	104.5	1614061
8	19900	115.7	7409732
9	8330000	65.3	8818928
10	12400	64.2	349223

Query 7: Does the aggregate continental urban population and total plastic waste show a trend of any kind?

As visualized by the correlation heatmap, I wanted to see how the high positive correlation in urban population and total plastic waste is demonstrated on the aggregate level. The table below ranks the urban population quantity of each continental region, with Asia having the highest total urban population and total plastic waste production. This kind of trend makes sense based on what we know about the regions. However what was surprising is that the Americas has a higher total population than Europe when the perception of Europe is having more urbanized land.



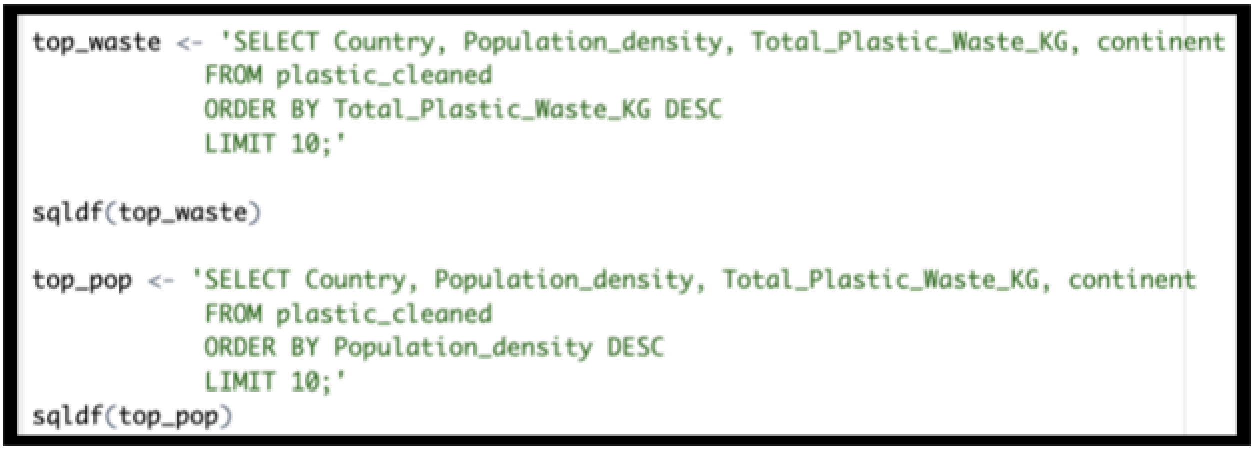
```
waste_urban <- 'SELECT continent,SUM(Urban_population) AS Total_Urban_Population,
                SUM(Total_Plastic_Waste_KG) AS Total_Plastic_Waste
                FROM plastic_cleaned
                GROUP BY continent
                ORDER BY Total_Plastic_Waste DESC;'

sqldf(waste_urban)
```



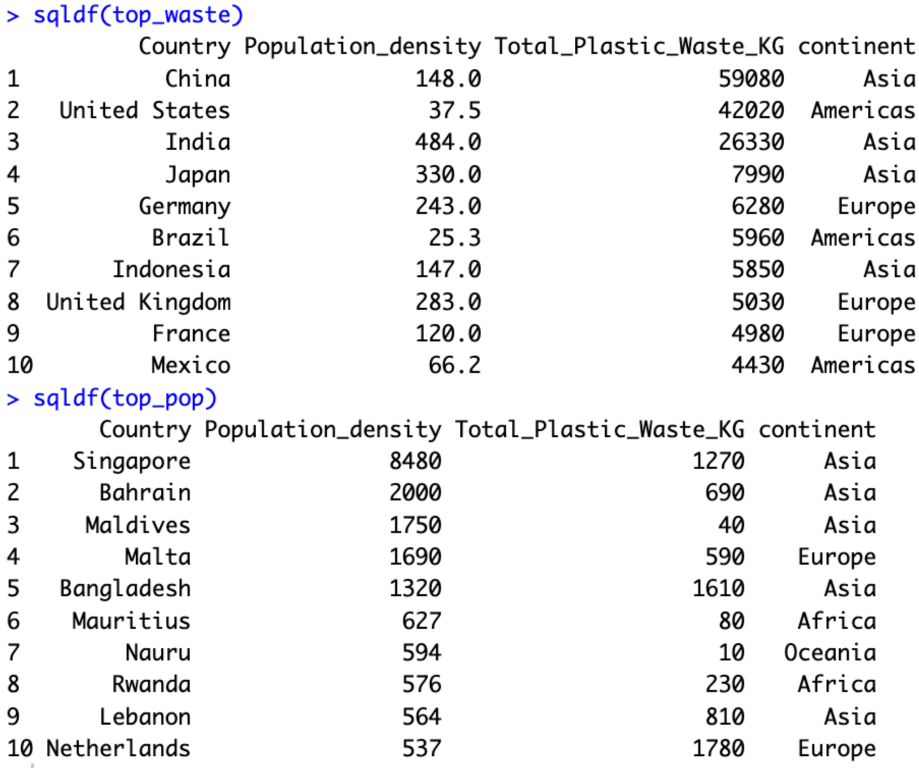
```
> sqldf(waste_urban)
  continent Total_Urban_Population Total_Plastic_Waste
1      Asia             2231115486              128270
2  Americas              783031479               69360
3    Europe              430599159               53580
4    Africa              586005324               12380
5   Oceania               25530457                2010
```

Query 8: Do the countries with the highest population density have the highest total plastic waste?

This query created two tables, one the top ten countries that have the highest total plastic waste amount and the other with the top ten countries with the highest population density. These tables were to understand whether there would be matching countries that have both the highest population density and total plastic waste amount. However, the two tables generated different countries when organized by descending order of either total plastic waste or population density. Within this query, we also chose to select continents to connect with the trend line visualization that is facet wrapped by continent. In the tables this only demonstrated that in both tables there were at least four out of the ten countries that were within the Asian continent. This lines up with our hypothesis because many Asian countries have high population density and also are known for manufacturing and exporting goods, which could contribute to the overall plastic waste production.

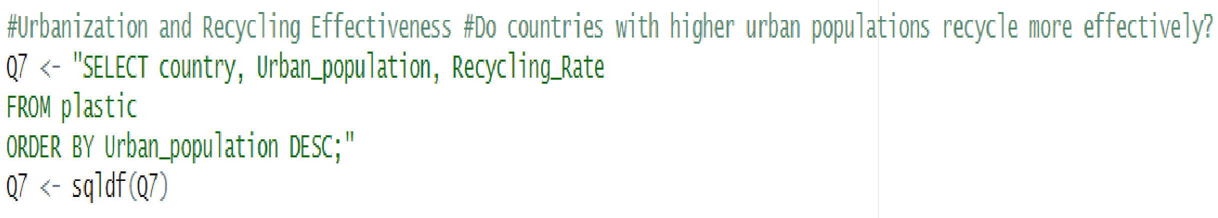


```
top_waste <- 'SELECT Country, Population_density, Total_Plastic_Waste_KG, continent
             FROM plastic_cleaned
             ORDER BY Total_Plastic_Waste_KG DESC
             LIMIT 10;'

sqldf(top_waste)

top_pop <- 'SELECT Country, Population_density, Total_Plastic_Waste_KG, continent
             FROM plastic_cleaned
             ORDER BY Population_density DESC
             LIMIT 10;'
sqldf(top_pop)
```



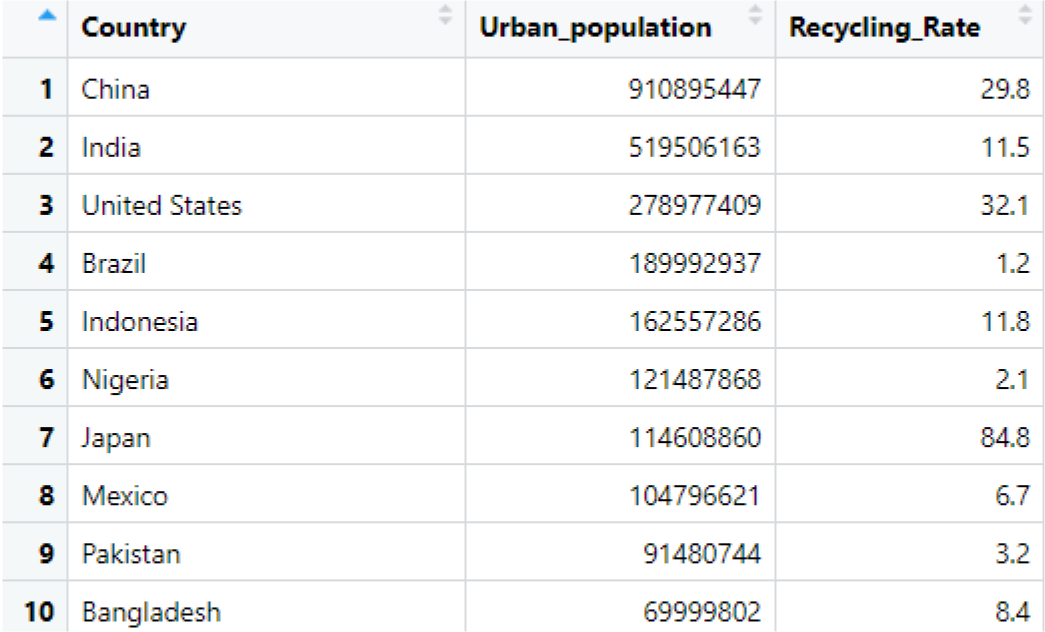
```
> sqldf(top_waste)
          Country Population_density Total_Plastic_Waste_KG continent
1           China              148.0                  59080      Asia
2   United States               37.5                  42020  Americas
3           India              484.0                  26330      Asia
4           Japan              330.0                   7990      Asia
5         Germany              243.0                   6280    Europe
6          Brazil               25.3                   5960  Americas
7       Indonesia              147.0                   5850      Asia
8  United Kingdom              283.0                   5030    Europe
9          France              120.0                   4980    Europe
10         Mexico               66.2                   4430  Americas
> sqldf(top_pop)
       Country Population_density Total_Plastic_Waste_KG continent
1    Singapore               8480                   1270      Asia
2      Bahrain               2000                    690      Asia
3     Maldives               1750                     40      Asia
4        Malta               1690                    590    Europe
5   Bangladesh               1320                   1610      Asia
6    Mauritius                627                     80    Africa
7        Nauru                594                     10   Oceania
8       Rwanda                576                    230    Africa
9      Lebanon                564                    810      Asia
10 Netherlands                537                   1780    Europe
```

Query 9: Do Countries with higher urban populations recycle more effectively?



```
#Urbanization and Recycling Effectiveness #Do countries with higher urban populations recycle more effectively?
Q7 <- "SELECT country, Urban_population, Recycling_Rate
FROM plastic
ORDER BY Urban_population DESC;"
Q7 <- sqldf(Q7)
```

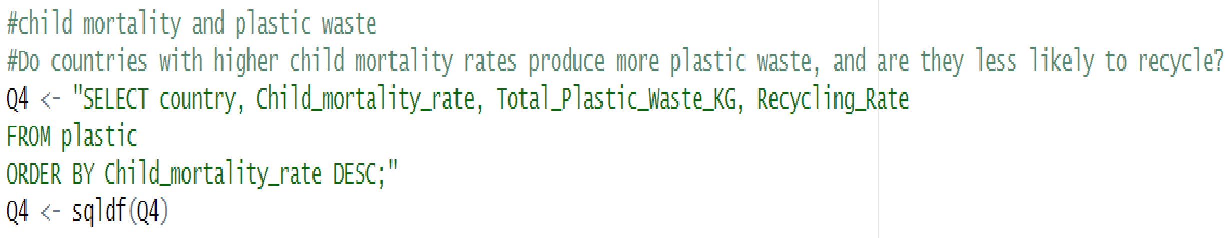
This visualization serves as a jumping off point to really see the impact of plastic waste beyond simple economy, it allows us to see if urbanization was truly a player in a possible solution to plastic waste or whether it was a main cause in the problem. There was no necessarily steady pattern between the variables but using outside knowledge looking at Japan which was the country with the highest rate of recycling we can gather that cultural rules and regulations if implemented strongly can definitely lead to a cleaner and greener environment.



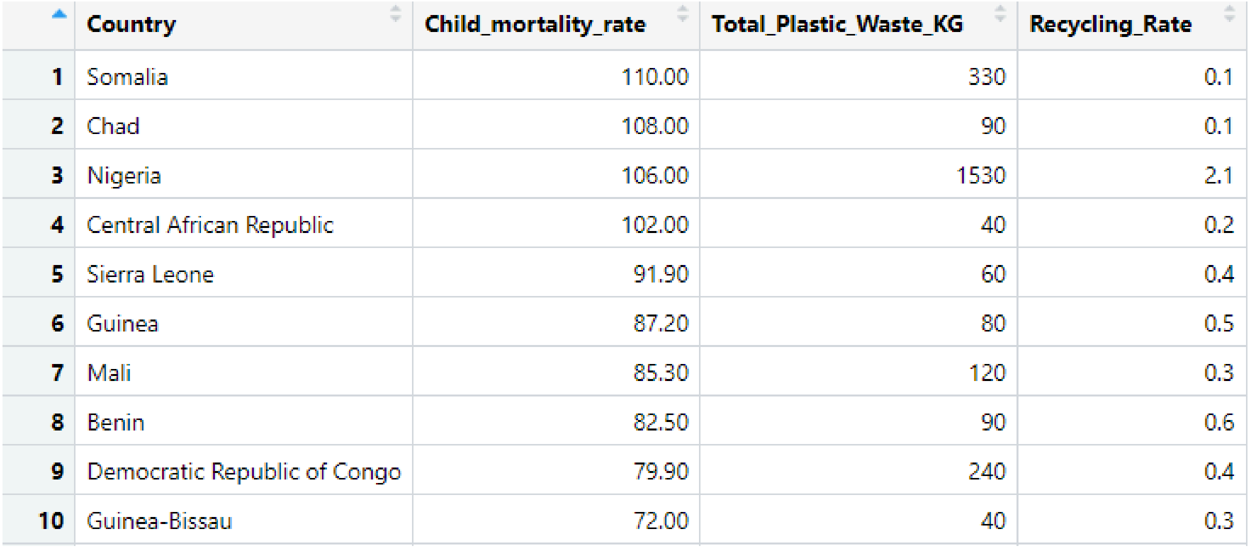
	Country	Urban_population	Recycling_Rate
1	China	910895447	29.8
2	India	519506163	11.5
3	United States	278977409	32.1
4	Brazil	189992937	1.2
5	Indonesia	162557286	11.8
6	Nigeria	121487868	2.1
7	Japan	114608860	84.8
8	Mexico	104796621	6.7
9	Pakistan	91480744	3.2
10	Bangladesh	69999802	8.4

Q10: Do Countries with Higher Child Mortality Rate Produce More Plastic Waste, and are they less likely to recycle? Pairs with visual 10.

In this small screen grab of the top ten countries there may not look like there is any terribly big news. However, when examining Visual 10, a clear pattern emerges. Countries with the highest peaks in child mortality are often represented by a very deep blue, indicating some of the lowest recycling rates. This suggests a significant lack of social commitment to clean up or mitigate environmental damage, which can have cascading negative effects on both the environment and children's well-being.



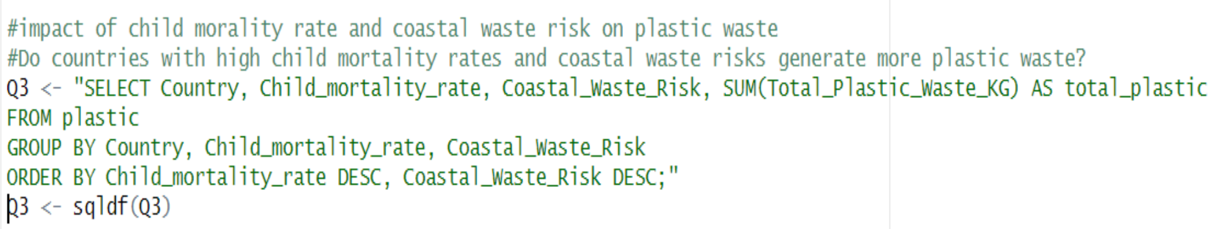
```
#child mortality and plastic waste
#Do countries with higher child mortality rates produce more plastic waste, and are they less likely to recycle?
Q4 <- "SELECT country, Child_mortality_rate, Total_Plastic_Waste_KG, Recycling_Rate
FROM plastic
ORDER BY Child_mortality_rate DESC;"
Q4 <- sqldf(Q4)
```



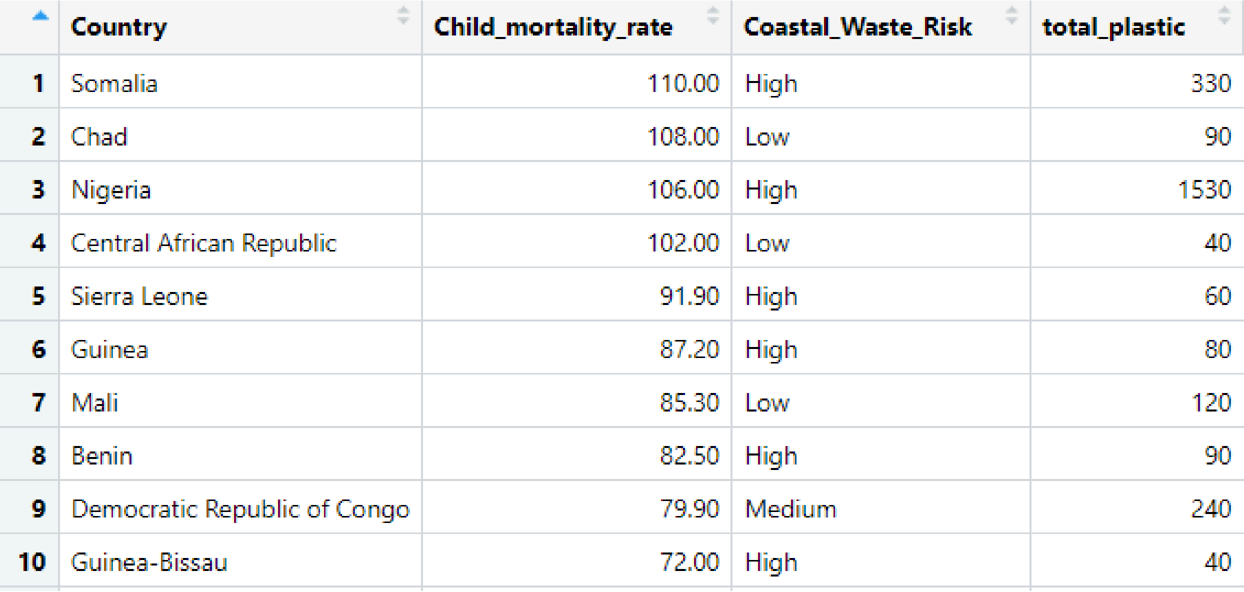
	Country	Child_mortality_rate	Total_Plastic_Waste_KG	Recycling_Rate
1	Somalia	110.00	330	0.1
2	Chad	108.00	90	0.1
3	Nigeria	106.00	1530	2.1
4	Central African Republic	102.00	40	0.2
5	Sierra Leone	91.90	60	0.4
6	Guinea	87.20	80	0.5
7	Mali	85.30	120	0.3
8	Benin	82.50	90	0.6
9	Democratic Republic of Congo	79.90	240	0.4
10	Guinea-Bissau	72.00	40	0.3

Q11: Do Countries with High Child Mortality Rates and Coastal Waste Risks generate more plastic waste?

This question explored how social factors, such as infant mortality rate, and environmental factors, such as coastal waste risk, are related to the production of plastic waste. At first glance, the visualization may not reveal much, as there is a fair amount of green distributed along each line. However, by focusing specifically on the 0-30 range of infant mortality rates within the “high” and “medium” coastal waste risk categories, a clearer pattern emerges. In these areas, although the infant mortality rate is relatively low, there is a significant risk of coastal waste accompanied by a high density of plastic waste. This highlights regions that may not yet suffer serious social consequences, but which are at substantial environmental risk due to poor management of plastic waste. This query provides valuable predictive information, underscoring the importance of monitoring these areas over time. By looking closely at these trends, policymakers and environmentalists can take early action to mitigate potential future social and environmental problems.



```
#impact of child morality rate and coastal waste risk on plastic waste
#Do countries with high child mortality rates and coastal waste risks generate more plastic waste?
Q3 <- "SELECT Country, Child_mortality_rate, Coastal_Waste_Risk, SUM(Total_Plastic_Waste_KG) AS total_plastic
FROM plastic
GROUP BY Country, Child_mortality_rate, Coastal_Waste_Risk
ORDER BY Child_mortality_rate DESC, Coastal_Waste_Risk DESC;"
Q3 <- sqldf(Q3)
```

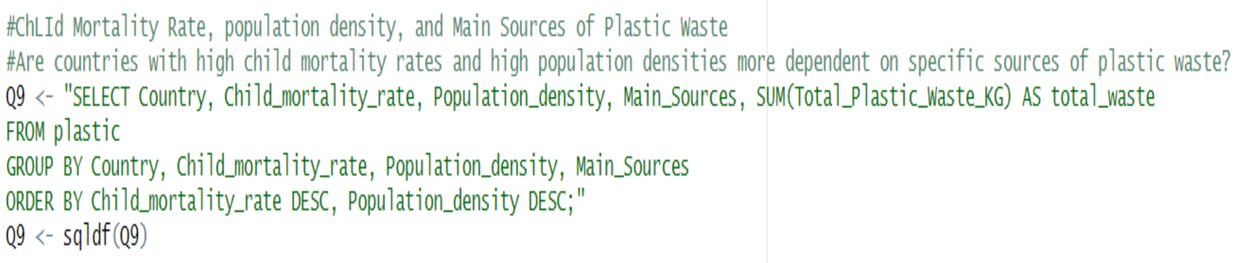


	Country	Child_mortality_rate	Coastal_Waste_Risk	total_plastic
1	Somalia	110.00	High	330
2	Chad	108.00	Low	90
3	Nigeria	106.00	High	1530
4	Central African Republic	102.00	Low	40
5	Sierra Leone	91.90	High	60
6	Guinea	87.20	High	80
7	Mali	85.30	Low	120
8	Benin	82.50	High	90
9	Democratic Republic of Congo	79.90	Medium	240
10	Guinea-Bissau	72.00	High	40

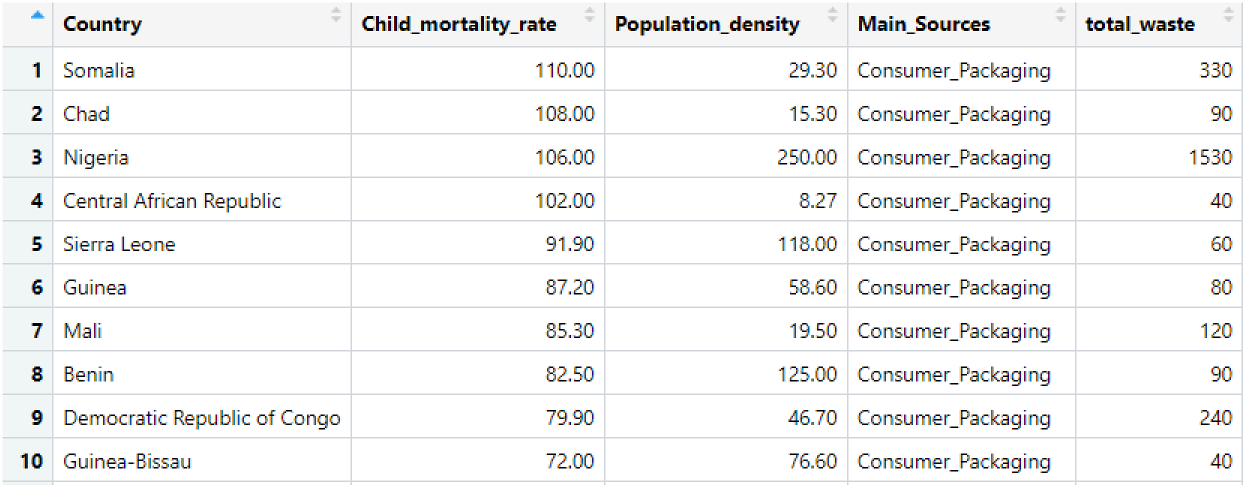
Q12: Are countries with high child mortality rates and high population densities more dependent on specific sources of plastic waste? Pairs with visual 12

This query and visualization represent the main correlation I wanted to explore throughout the process. I had read a lot about the positive correlation between air pollution and stillbirth rates, which sparked my interest in investigating whether plastic waste could have similar effects. When analyzing the data, I focused on the relationship between child mortality and the dominant categories of plastic waste.

The results showed that in countries with the highest child mortality rates, consumer packaging is the dominant category of plastic waste. This discovery particularly intrigued me. While correlation doesn't equal causation, it led me to theorize about the potential causes of this relationship. Consumer packaging is a broad category that includes items such as clothing tags, candy wrappers and fruit stickers. All this may contribute to the formation of microplastics, which are increasingly detected in the placentas of pregnant mothers. Depending on the chemicals used in the production of these packaging materials, adverse effects on fetal development may occur. Although further research is needed to investigate this potential link, these findings highlight an area of serious concern that requires closer examination.



```
#ChLId Mortality Rate, population density, and Main Sources of Plastic Waste
#Are countries with high child mortality rates and high population densities more dependent on specific sources of plastic waste?
Q9 <- "SELECT Country, Child_mortality_rate, Population_density, Main_Sources, SUM(Total_Plastic_Waste_KG) AS total_waste
FROM plastic
GROUP BY Country, Child_mortality_rate, Population_density, Main_Sources
ORDER BY Child_mortality_rate DESC, Population_density DESC;"
Q9 <- sqldf(Q9)
```



	Country	Child_mortality_rate	Population_density	Main_Sources	total_waste
1	Somalia	110.00	29.30	Consumer_Packaging	330
2	Chad	108.00	15.30	Consumer_Packaging	90
3	Nigeria	106.00	250.00	Consumer_Packaging	1530
4	Central African Republic	102.00	8.27	Consumer_Packaging	40
5	Sierra Leone	91.90	118.00	Consumer_Packaging	60
6	Guinea	87.20	58.60	Consumer_Packaging	80
7	Mali	85.30	19.50	Consumer_Packaging	120
8	Benin	82.50	125.00	Consumer_Packaging	90
9	Democratic Republic of Congo	79.90	46.70	Consumer_Packaging	240
10	Guinea-Bissau	72.00	76.60	Consumer_Packaging	40

Conclusion

Over the course of this project, we conducted various visualizations and queries analyzing the data we had gathered. Through this process we encountered multiple hurdles in the readability of visualization which emphasized the importance of scaling data. Also the importance of the observational distribution, most of the filtering of the data was based on matching data for the countries we had at our disposal without regard to ensuring that there was an equal distribution of countries represented from each continental region. Looking back at how we would approach this differently would be to ensure that there is equal representation of each continental region so that it doesn't skew the data analysis we conduct. Overall working within the time constraint of this project, we weren't able to do the extensive research of policy or creating more variables to account for more characteristics of the data.

Regardless of the hurdles we faced, there were still some valuable insights gained from this analysis. Something to account for was the fact that we could have included more variables such as other welfare indicators, geographical data, and regulation data. However, based on the data we analyzed, the biggest take away from these is that regardless of economic power represented by each country globally there isn’t enough being done to mitigate this environmental issue that has real effects on the global population. Also with consumer packaging being a highlighted source of plastic waste that puts some responsibility into the hands of consumers, as a reminder that there are individual behavioral changes that can be made to make a difference. Also this could hopefully be used to highlight that there are industries that can be more closely regulated by policy to help mitigate some of this plastic waste production.