FA582 Assignment 1 Report

Naveen Mathews Renji - 20016323

Problem 1

Use the datasets provided for Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Do the following:

- Load in and clean the data.
- Conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc.
- Conduct exploratory data analysis to visualize and make comparisons for residential building category classes across boroughs and across time (select the following: 1-, 2-, and 3-family homes, coops, and condos). Use histograms, box plots, scatterplots or other visual graphs. Provide summary statistics along with your conclusions.

1. Data Loading and Transformation (Handling missing values, outliers and zero values)

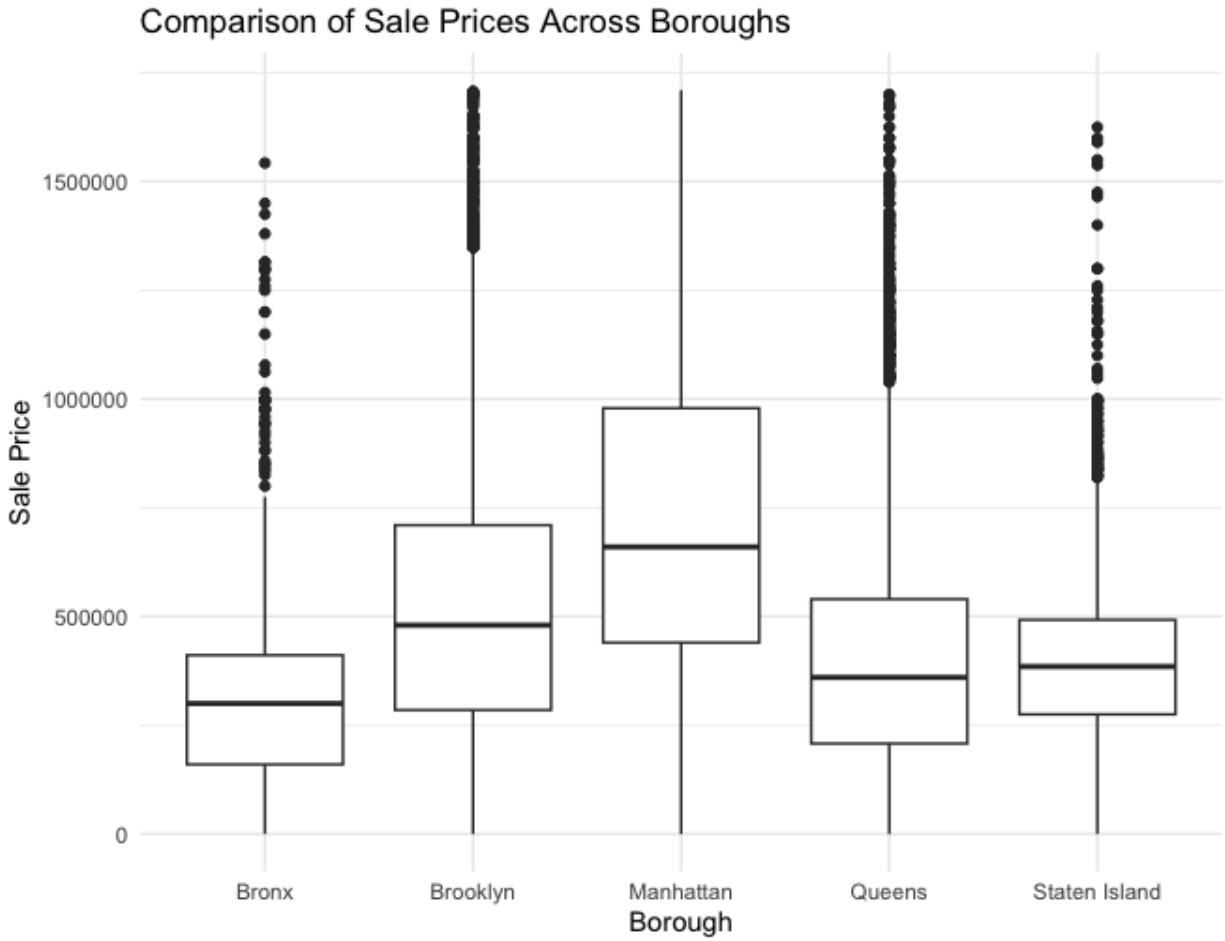
1. I loaded the 5 different excel files, then added a column to each of them called borough which was filled with their respective borough names. Then I concatenated the 5 dataframes into one.
2. All column names are converted to lowercase to maintain consistency and facilitate easy referencing.
3. The sale date in the dataset is converted from Excel date format to a more human-readable date-time format
4. I checked for missing values using sapply() and found that there are no missing values in any of the columns.
5. A total of 28,638 records have a sale price of zero, which are not useful for our analysis and are therefore removed.
6. Using the IQR method, I identified 6,309 outliers in the sale.price column, which were then removed.
7. I also addressed the issue of zero values in the year.built column by replacing them with the median value.

2. Performing EDA

- Created a new dataframe with only these building categories

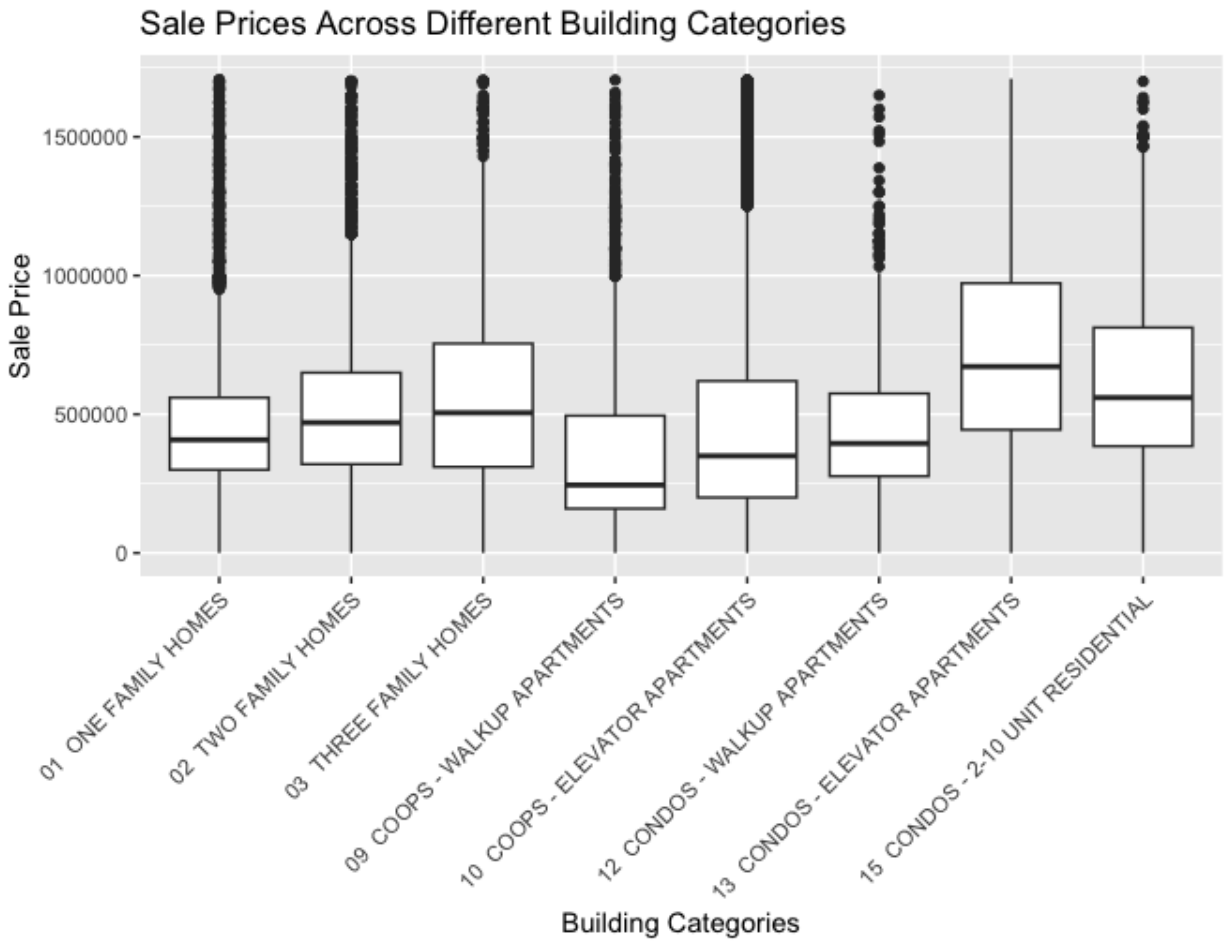
```
categories_to_consider <- c(
  '01  ONE FAMILY HOMES', '02  TWO FAMILY HOMES', '03  THREE FAMILY HOMES',
  '09  COOPS - WALKUP APARTMENTS', '10  COOPS - ELEVATOR APARTMENTS',
  '12  CONDOS - WALKUP APARTMENTS', '13  CONDOS - ELEVATOR
APARTMENTS',
  '15  CONDOS - 2-10 UNIT RESIDENTIAL'
)
```

- Comparison of Sale Prices Across Boroughs



This plot gives us an idea of the sale prices varying across the various boroughs. I observed that Manhattan had the most expensive sales followed by Brookyn, Queens, Staten Island and finally the cheapest being the Bronx

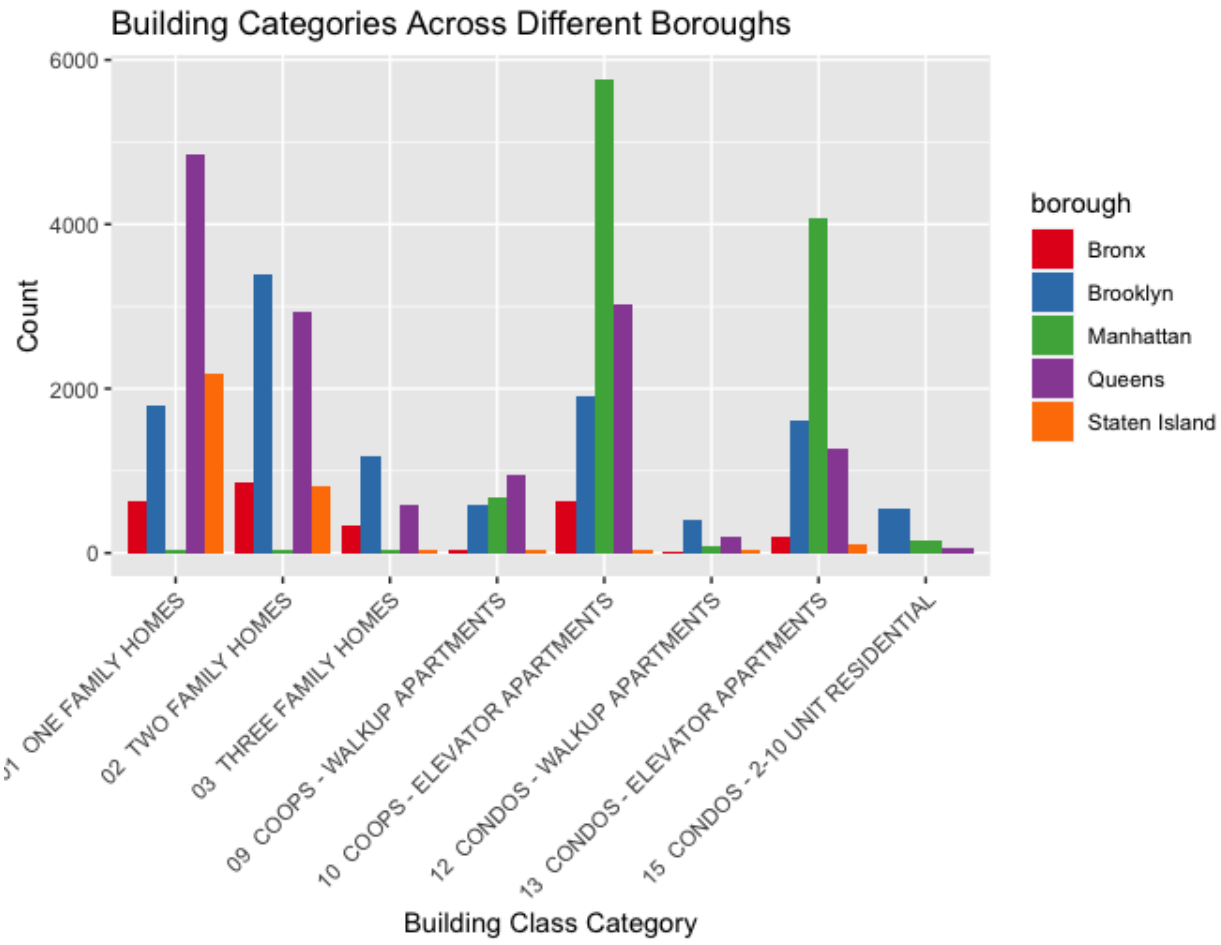
- Sale Prices Across Different Building Categories.



From this plot, we are able to visualize and analyze how the type of unit or the category of the building plays a vital role in determining the sale price.
The prices indicate that

1. Walkup apartments are more likely to be cheaper.
2. Family homes show a similarity in range of prices with all three categories being available in the similar sale price but the maximum cost differs across boroughs.
3. Coops are the most affordable real estate in the New York Boroughs.
4. Condos are on the more expensive end with elevator apartment condos being the most expensive of them all across the categories that we are considering.

- Building Categories Across Different Boroughs



Through this plot, we are able to visualize how the different boroughs portrays itself to cater to a different crowd altogether.

- If we were to draw a conclusion on family living from this plot, it would be that most homes in Queens are family homes and CO-OP Apartments, although 3 family homes are not very common as compared to Brooklyn. Queens too has a good share of sales in condos and coops.

- The next conclusion we can draw from this plot is that most residential units in Manhattan that are sold are ones belonging to elevator apartments both Co-ops and Condos, a small percentage of walkup Coop apartments and almost a negligible share for the rest of the residential categories.

- Brooklyn appears to be the most diverse residential borough that caters to a variety of individuals and groups and income categories.

- Staten Island majorly attracts only families and the residences seem to be mostly single units family homes.

- The Bronx seems to be attracting the least number of buyers indicating a problem that needs to be solved to attract more homeowners. These could be social and local factors such as crime rate, lack of quality educational institutions for children and healthcare.

- Sale Price across Boroughs over Time

This graph shows an almost straight line for the sale prices across the year 2012 to 2013 indicating very little change in the sale prices or real estate value of residential homes that year. The only exception being Brooklyn that has shown a slight increase in the sale prices indicating a positive real estate value of residences in the Brooklyn Borough.

Problem 2

- The datasets provided nyt1.csv, nyt2.csv, and nyt3.csv represent three (simulated) days of ads shown and clicks recorded on the New York Times homepage. Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number impressions, number of clicks, and logged-in. Use R to handle this data. Perform some exploratory data analysis:
    - Create a new variable, age_group, that categorizes users as “<20”, “20-29”, “30-39”, “40-49”, “50-59”, “60-69”, and “70+”.
    - For each day:
        - Plot the distribution of number of impressions and click-through-rate (CTR = #clicks / #impressions) for these age categories
        - Define a new variable to segment or categorize users based on their click behavior.
        - Explore the data and make visual and quantitative comparisons across user segments/demographics (<20-year-old males versus <20-year old females or logged-in versus not, for example).
    - Extend your analysis across days. Visualize some metrics and distributions over time.

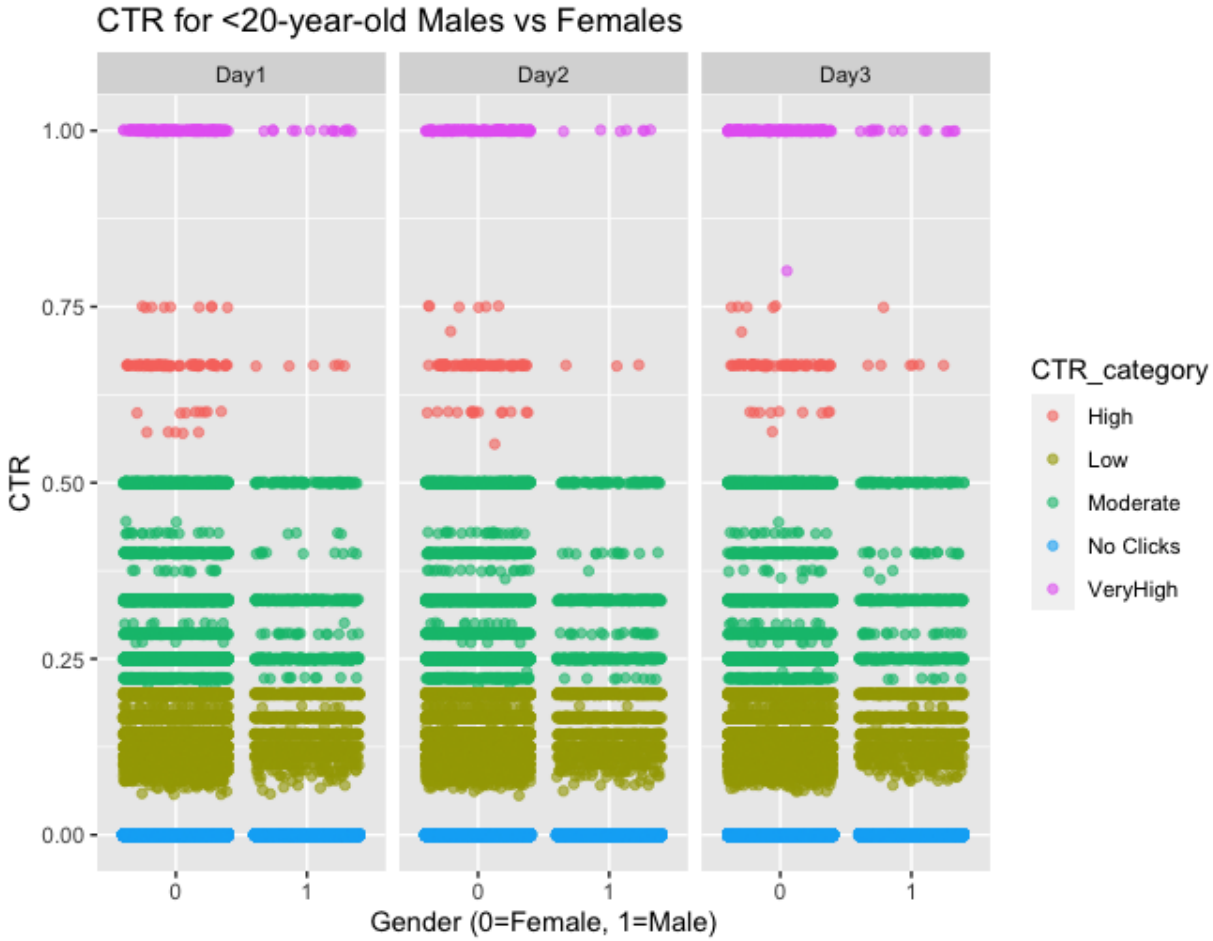
1. Data Loading and Transformation (Handling missing values, outliers and zero values)

- Data Loading
    - Three separate datasets named nyt1.csv, nyt2.csv, and nyt3.csv were loaded into the R environment. Each dataset represents a simulated day of ads shown and clicks recorded on the New York Times homepage. After loading, a new column Day was added to each dataset to signify which day the data belongs to:
        - nyt1 data was tagged with Day1
        - nyt2 data was tagged with Day2
        - nyt3 data was tagged with Day3
- The three separate datasets were then combined into a single dataset named nyt_data using the rbind function.
- The following steps were taken to clean the data:
    - Filtering: Rows where the Impressions column had a value of zero were removed.
    - Type Conversion: The Gender column was converted to a factor variable.
    - New Variables:

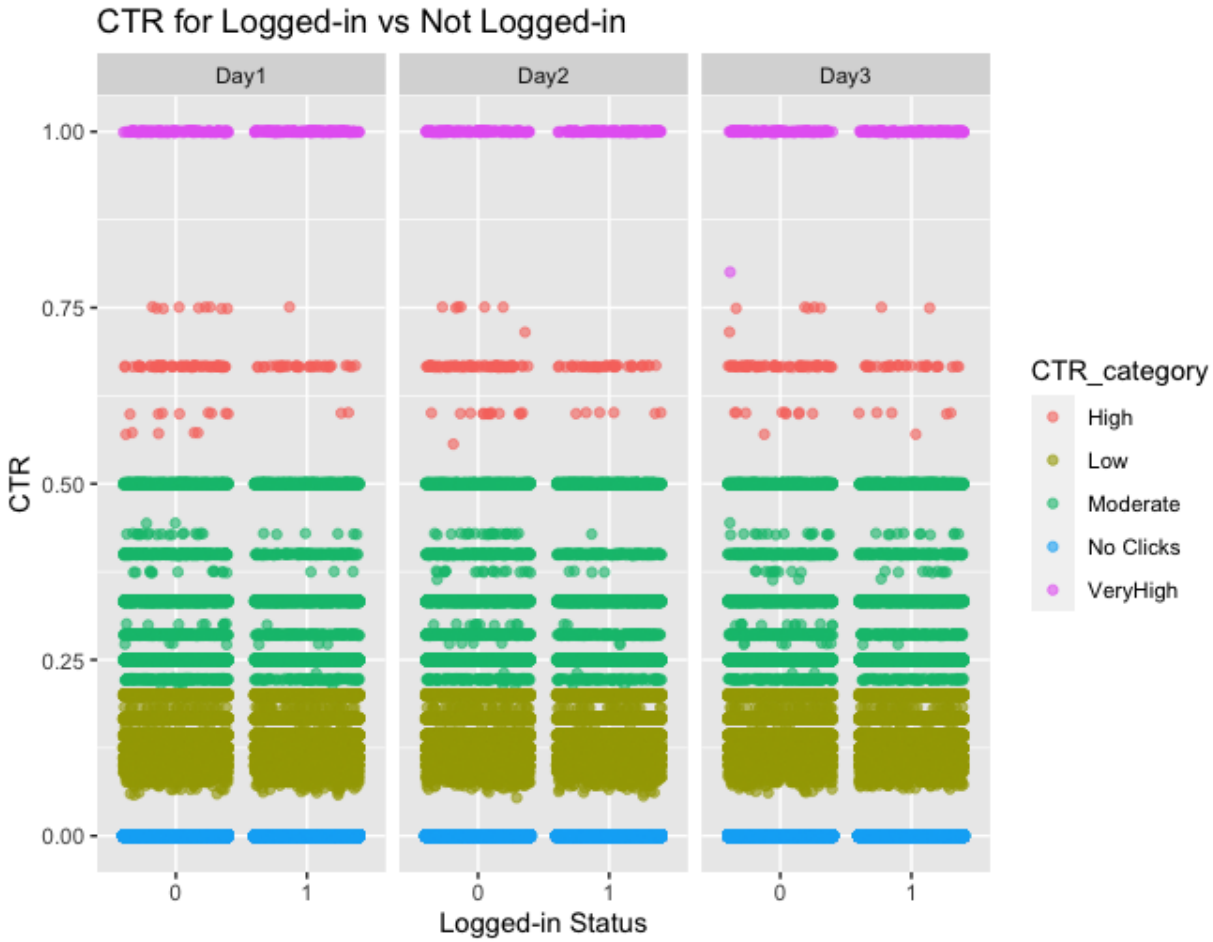
        - age_group: A new variable was created to categorize users based on their age into groups such as <20, 20-29, and so on.
        - CTR: Click-through rate (CTR) was calculated as # clicks / # impressions.
    - Handling Outliers and Missing Data:
        - A median age was calculated, only considering ages between 1 and 90.
        - If an Age was less than 1 or greater than 90, it was replaced with this median age.
    - Click Behavior Categorization:
        - A new variable click_category was created to categorize users based on whether they clicked an ad, didn't click, or had no impressions.

2. Performing EDA

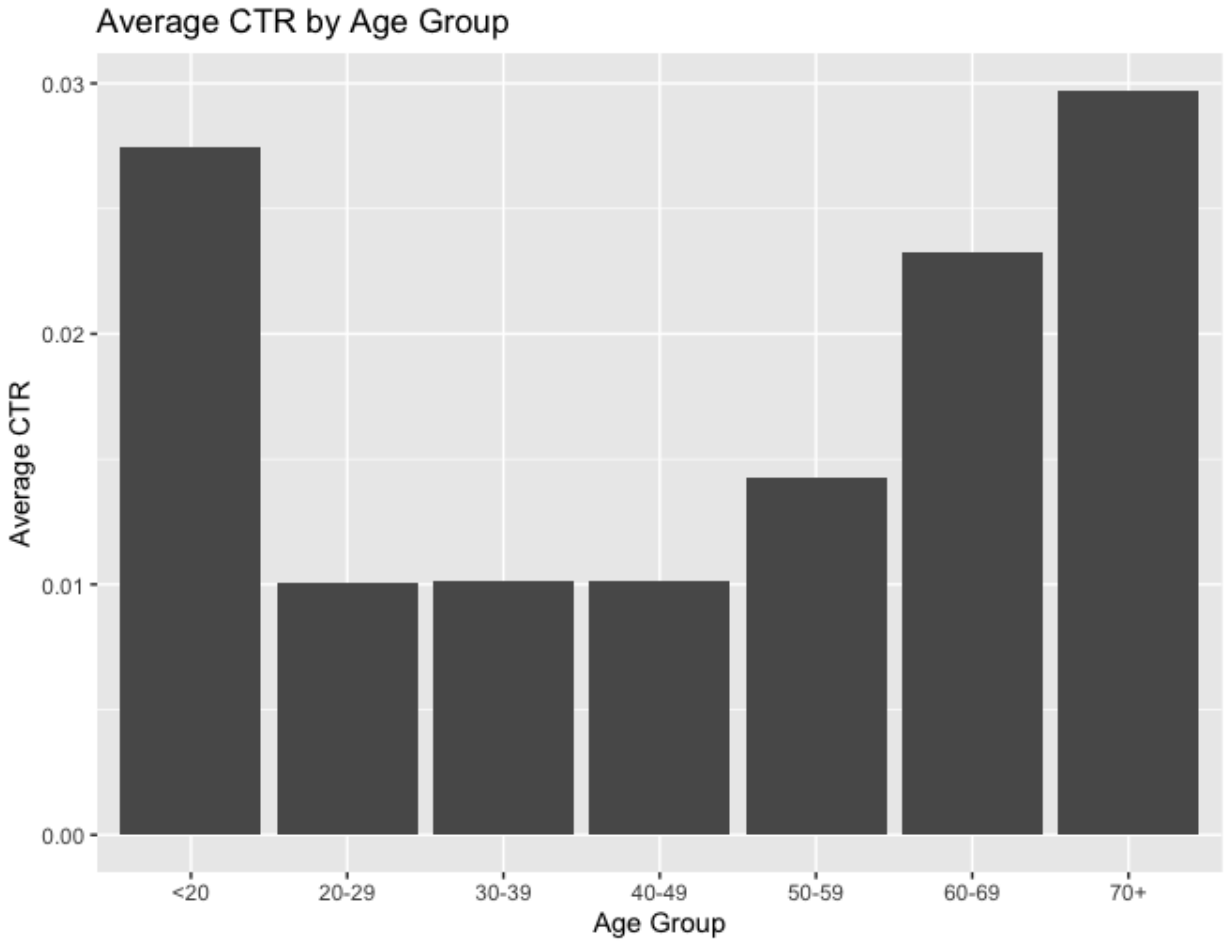
- "CTR for <20-year-old Males vs Females" - Compares click-through rates for young males and females.



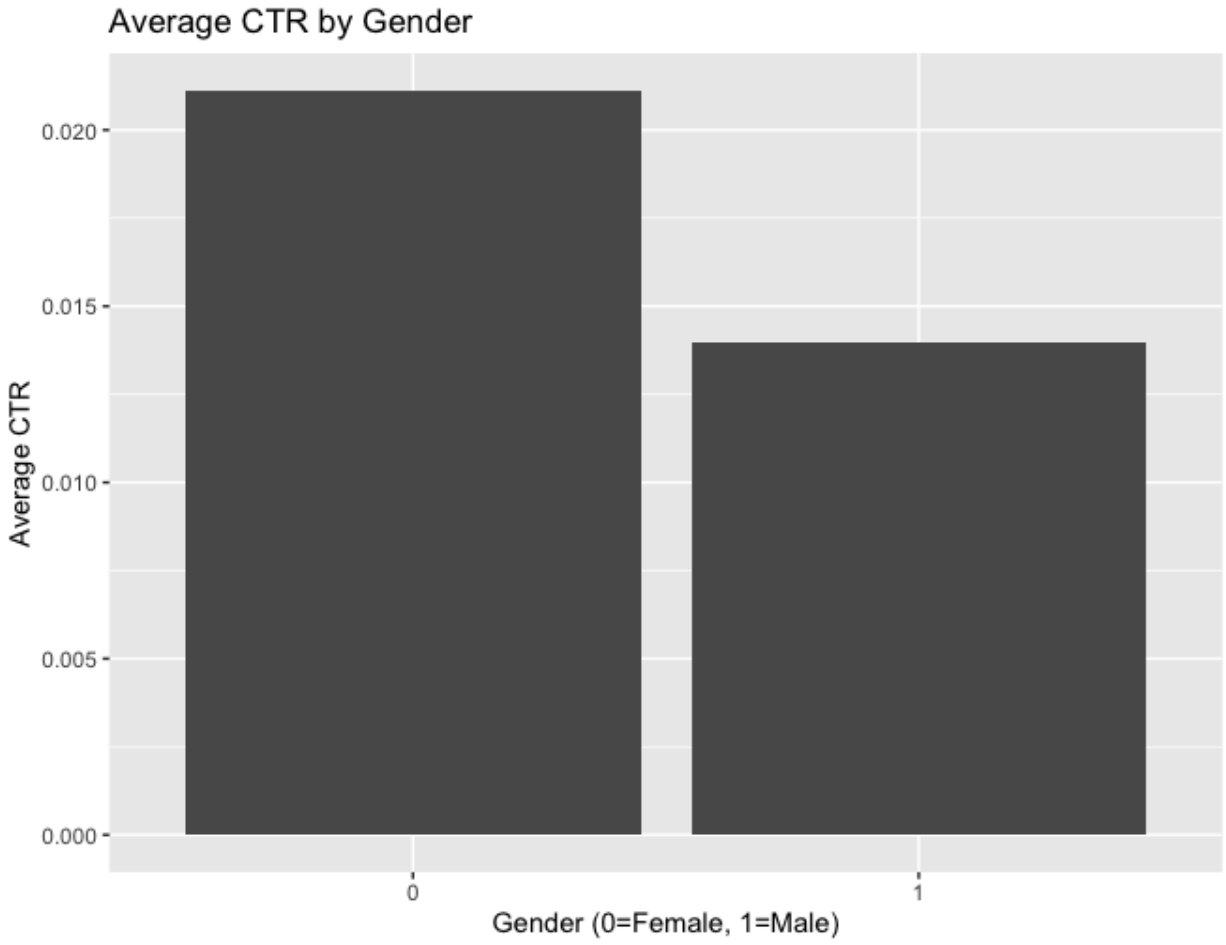
- "CTR for Logged-in vs Not Logged-in" - Compares click-through rates for logged-in and non-logged-in users.



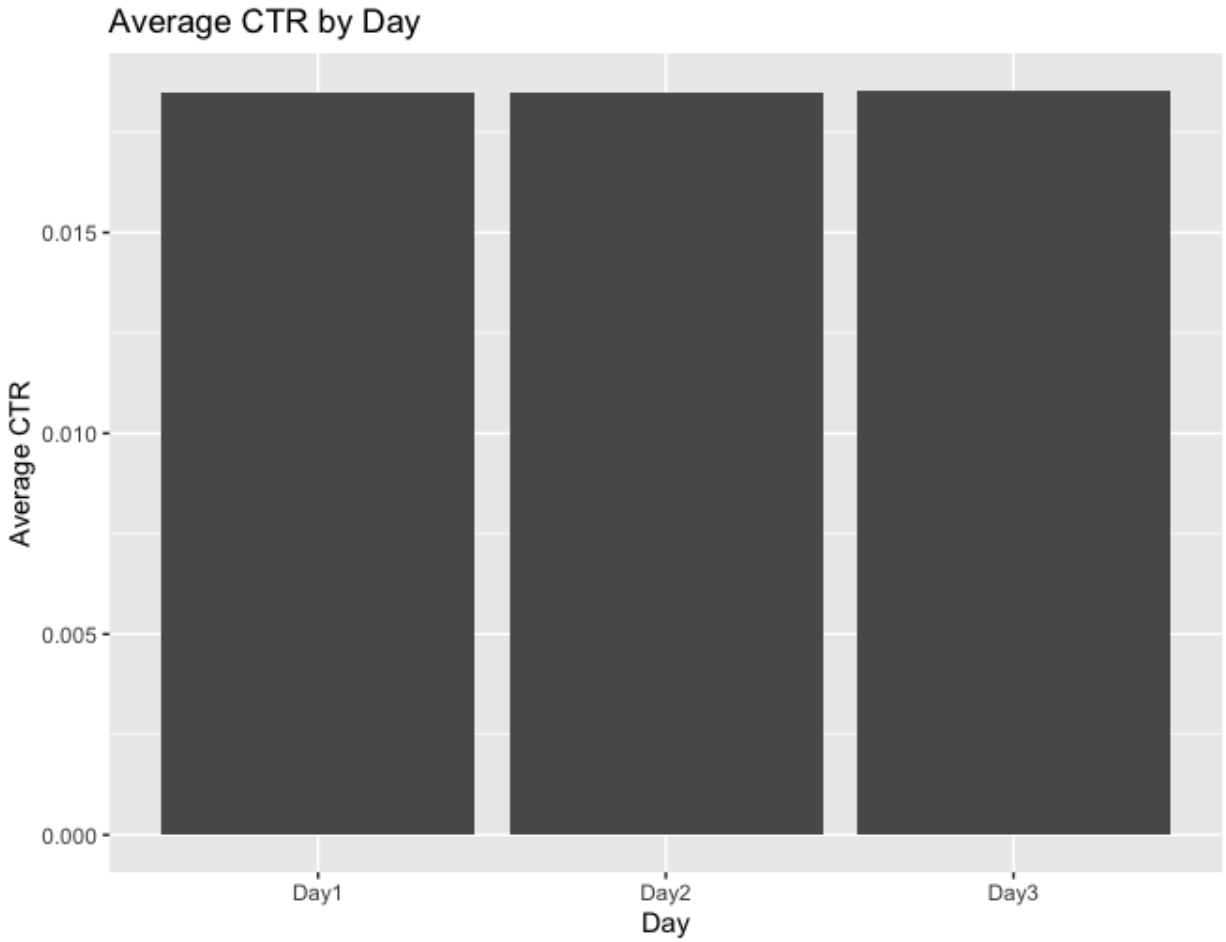
- "Average CTR by Age Group" - Shows average click-through rates for different age groups.



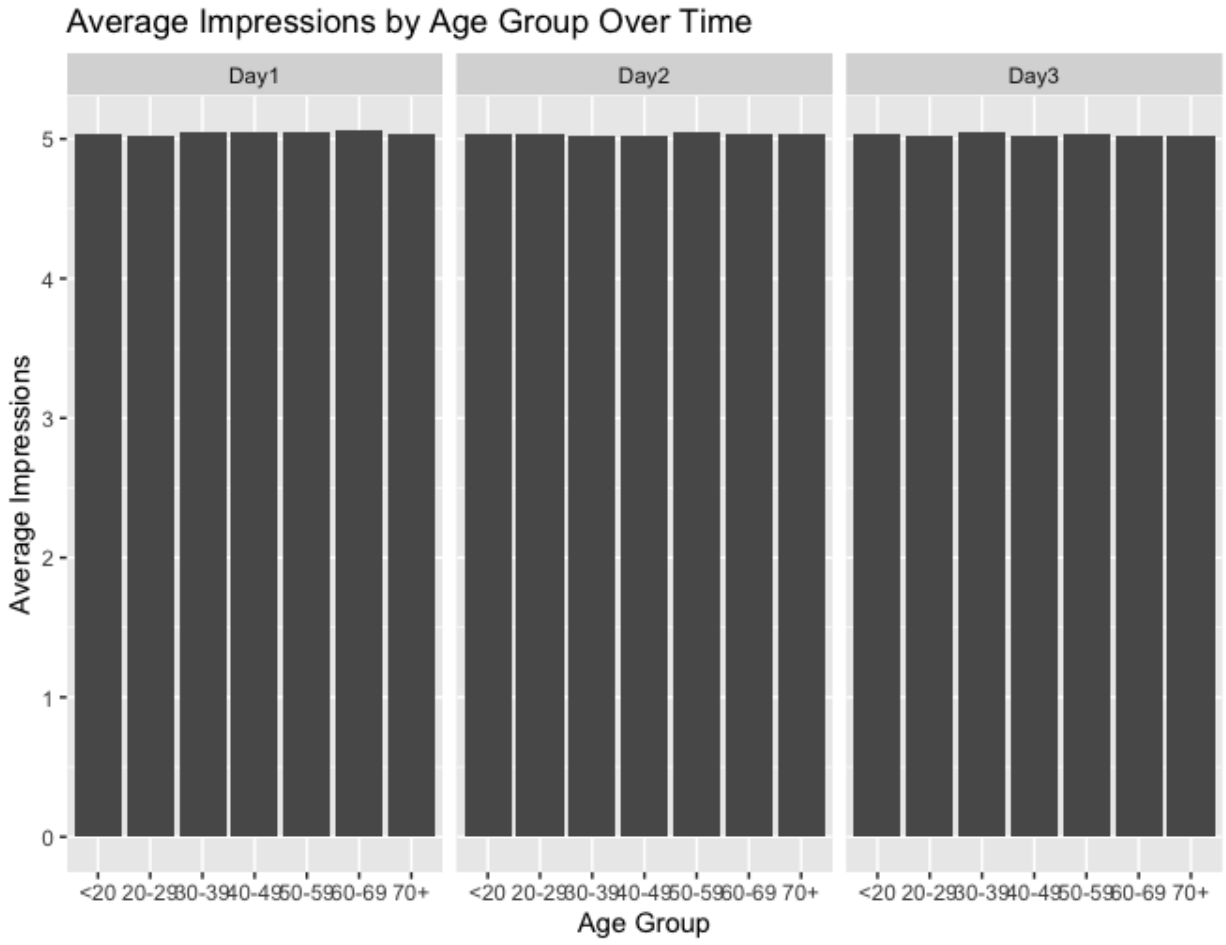
- "Average CTR by Gender" - Shows average click-through rates for males and females.



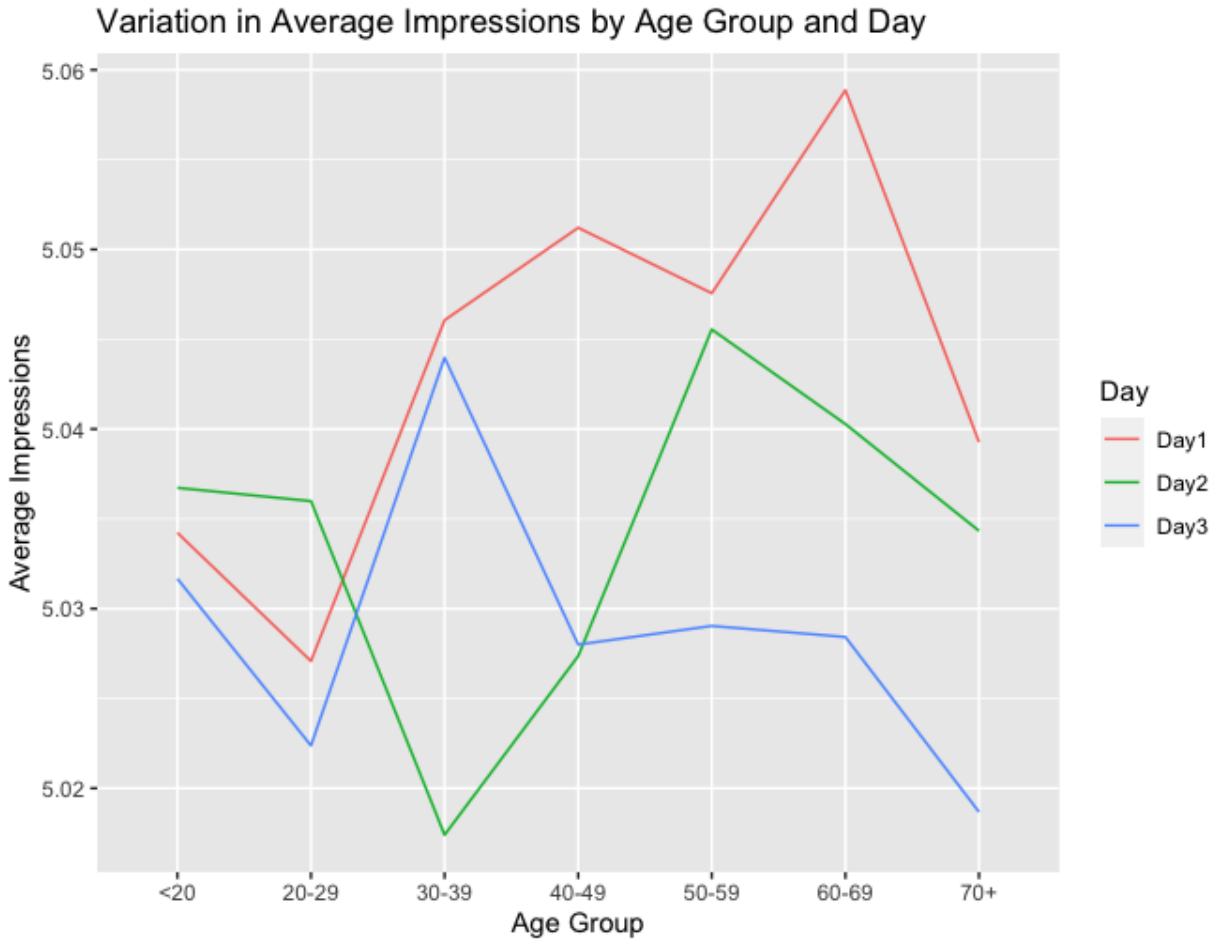
- "Average CTR by Day" - Shows average click-through rates for each day.



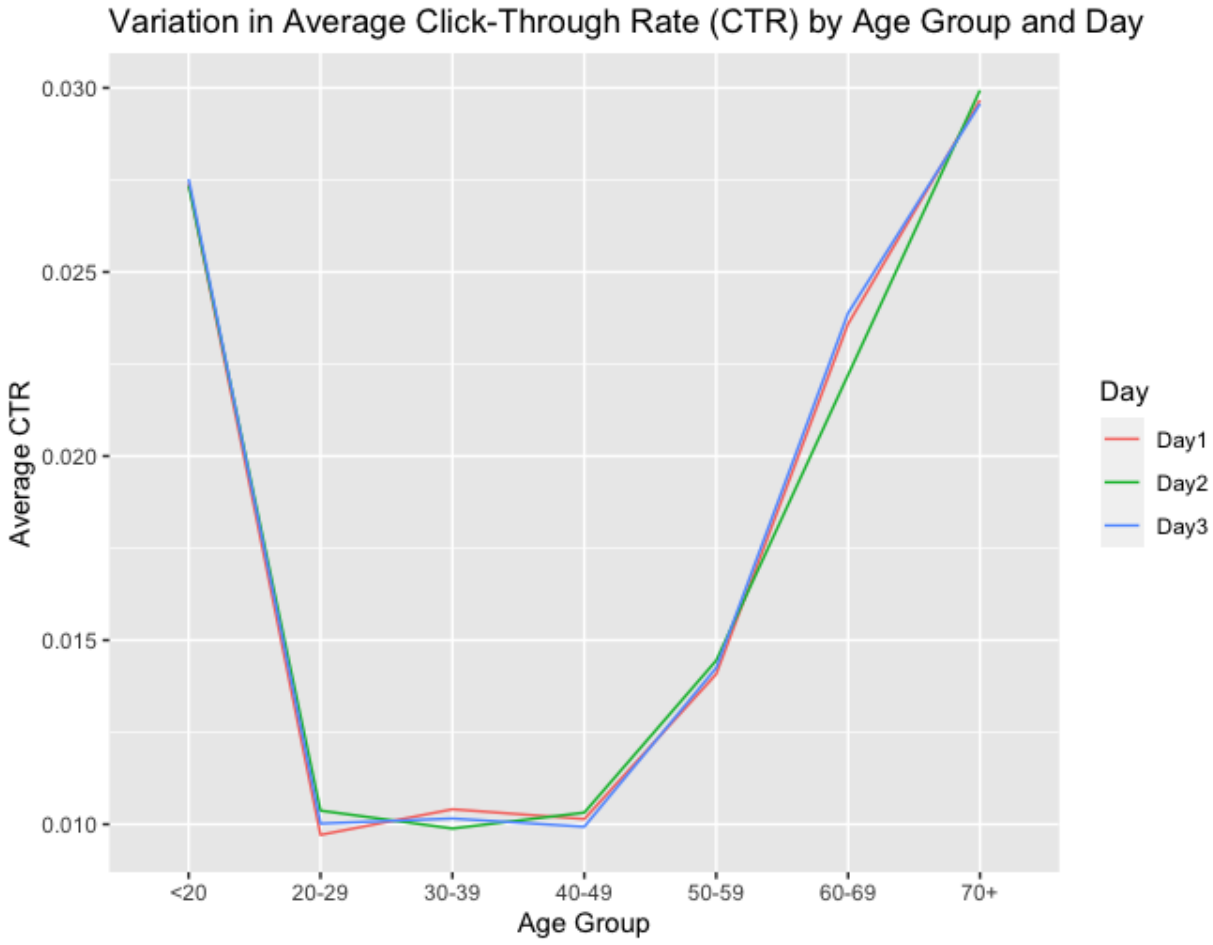
- "Average Impressions by Age Group Over Time" - Shows average impressions for different age groups over days.



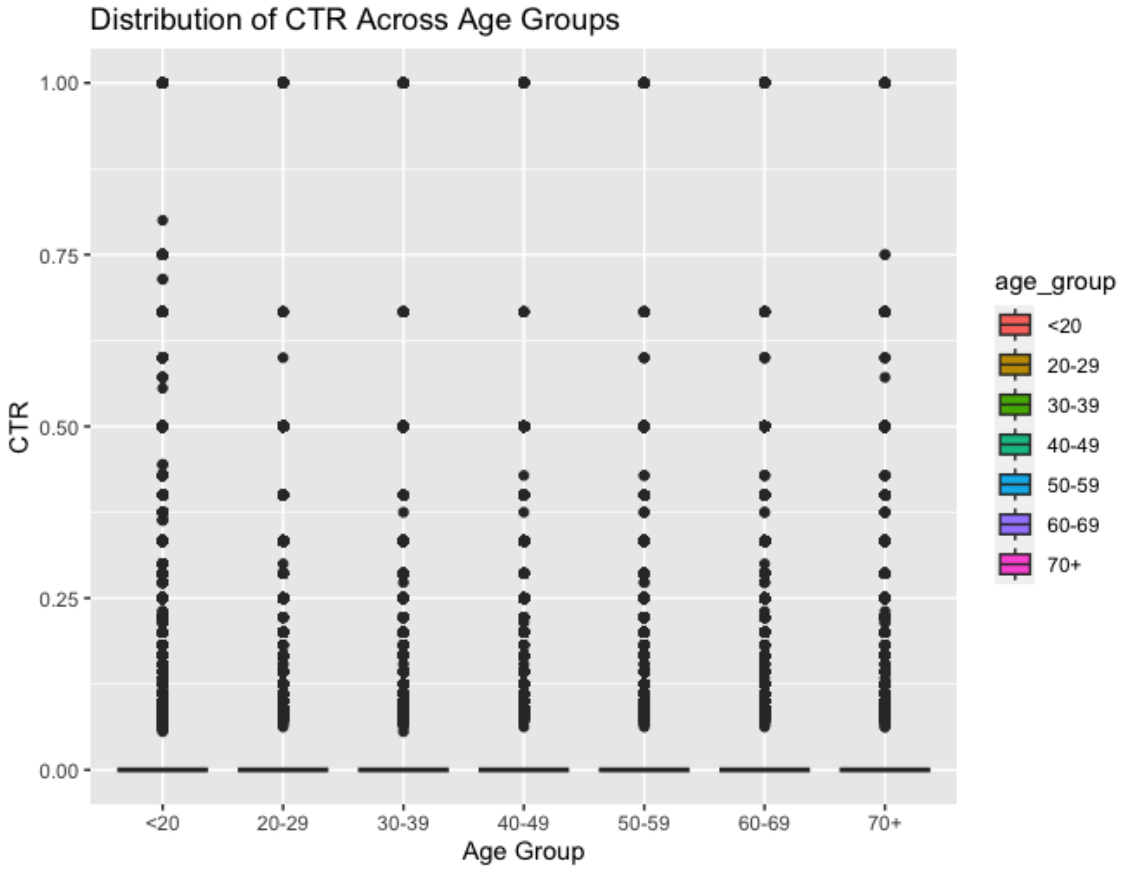
- "Variation in Average Impressions by Age Group and Day" - Line plot for average impressions by age and day.



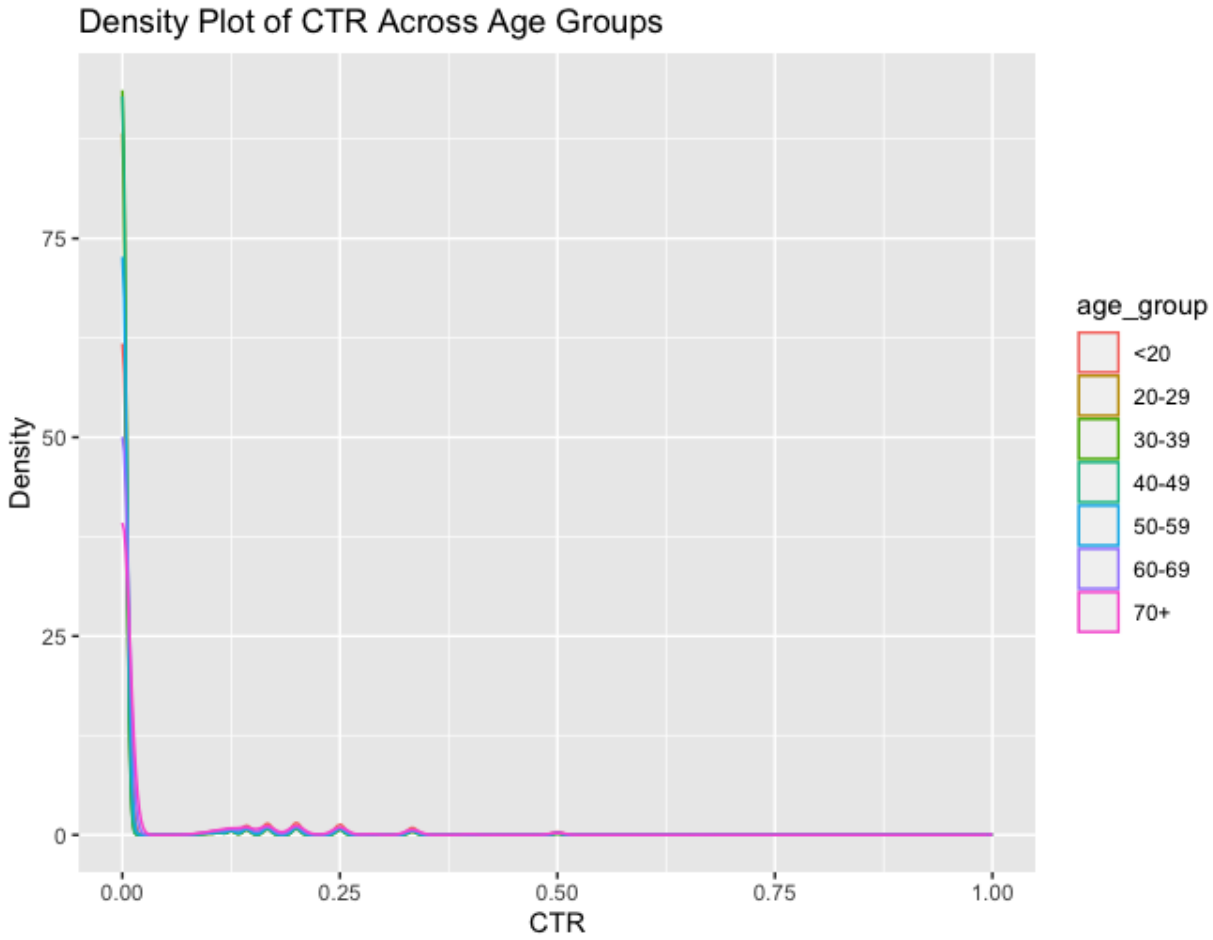
- "Variation in Average Click-Through Rate (CTR) by Age Group and Day" - Line plot for average CTR by age and day.



- "Distribution of CTR Across Age Groups" - Boxplot showing the distribution of CTR across age groups.



- "Density Plot of CTR Across Age Groups" - Density plot for CTR distribution across age groups.



3. Conclusion

After an in-depth analysis of the New York Times ad click data, several key observations have been made regarding Click-Through Rates (CTR), gender, and age groups.

Overall CTR Trends
The average CTR was relatively constant over the span of three days, hovering around 0.01847. This suggests that the engagement level with the ads was stable over this period.

Gender-Based CTR
The data indicated a noticeable difference in CTR between males and females. On average, females showed a higher CTR of approximately 0.0211 compared to 0.0140 for males. This suggests that females were more likely to engage with the advertisements.

CTR Across Age Groups
Analysis of different age groups highlighted some interesting trends:

- For the age groups between 20-50, the CTR remained consistent, around 0.01.
- The younger age group (<20) showed a significantly higher CTR of about 0.027.
- The older age groups, specifically 50-59, 60-69, and 70+, showed CTRs of around 0.014, 0.023, and 0.029 respectively.

Day-wise Analysis
When breaking down the data by days, it was found that mean impressions for all age groups remained fairly constant at approximately 5 impressions. However, for age groups over 50, there was a reduction in impressions after day 1, but the CTR remained consistent.

Overall, the analysis provides valuable insights into user engagement with New York Times advertisements. Female users and users either younger than 20 or older than 50 are more likely to click on the ads. These findings could be instrumental for targeted advertising strategies.

The consistency in CTR across the three days also offers some confidence in the reliability of these observed trends, making them potentially useful for long-term advertising planning