Final Project Report
Group 21

Analyzing: Who Gets The Final Rose?

Charlotte Parent, Jing He, Lynn Lam, Maya Rai

SI 370 | November 2024

Introduction

The Bachelor/Bachelorette series, which debuted in 2003 and has since captivated millions of viewers, has become a cultural phenomenon in reality television. Beyond its entertainment value, the show's success offers significant data-driven insights. Our project aims to uncover trends that can inform strategic business decisions by examining contestant characteristics, audience preferences, and engagement metrics. These insights could benefit networks, advertisers, and contestants alike.

This project report details our data analysis process, findings, and actionable recommendations, focusing on maximizing audience engagement and optimizing the show's success.

Project Goals

Our project aims to analyze data from the Bachelor/Bachelorette TV series to uncover trends, insights, and patterns that could be applied to other shows seeking to replicate the series's success. Specifically, our analysis seeks to:

1. Determine which version of the show (Bachelor or Bachelorette) has better ratings
2. Analyze if the age groups of the contestants are similar between the shows
3. Investigate the geographical patterns in the contestants' hometowns and if there were any shifts over the seasons
4. Examine factors/features impacted by the show, such as the age and job of the contestants

By doing so, we hope to understand contestant characteristics that correlate with success on the show, identify viewer preferences and potential factors that drive engagement, and explore how the findings could be leveraged for marketing strategies, such as targeted advertising or content creation. By aligning insights from our analysis with business objectives, the results from the data could help networks optimize programming and sponsorship decisions or assist contestants in strategizing their public personas.

Data Cleaning & Pre-Processing

Data Source

Kaggle Link

The data includes contestant information (demographics, occupations, and elimination details), episode summaries (engagement metrics like viewer and rating details), and seasons and specials (historical and spin-off details). Our analysis focused on contestants and episodes.

We focused on key variables like:

1. Demographic Information (age, hometown, and occupation)
2. Engagement metrics (viewers per episode)
3. Show dynamics (screen time and episode longevity)

Data Processing Steps Taken

1. **Missing Values:** Handled missing or inconsistent values, such as contestants without age data, dropping rows with NaNs given the small amounts of rows missing
2. **Formatting:** Standardized column names and ensured consistent formatting for date and text fields, such as the ratings column, converting date columns to datetime, renaming columns like 'No Overall' to 'Number Overall," removing unwanted columns (e.g., 'Prod Code' from the episodes dataset), etc.
3. **Feature Engineering:** Changed Ratings column to have a consistent representation, and Production Code was dropped as an unhelpful column

Exploratory Data Analysis (EDA)

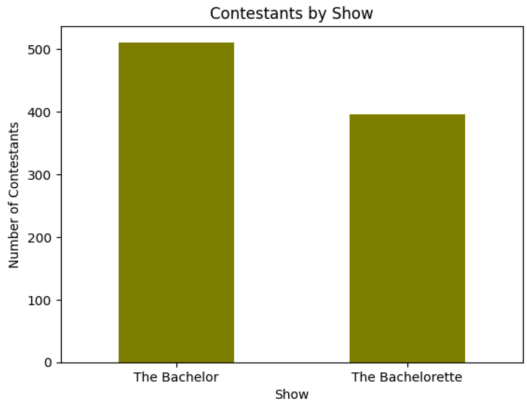
<u>Contestants EDA</u>

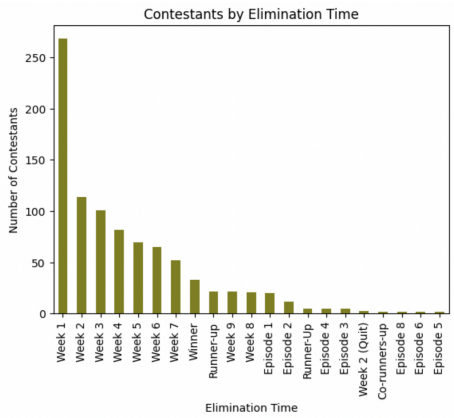
Descriptive Statistics:

1. Average contestant age: 27
2. Number of unique jobs: 610

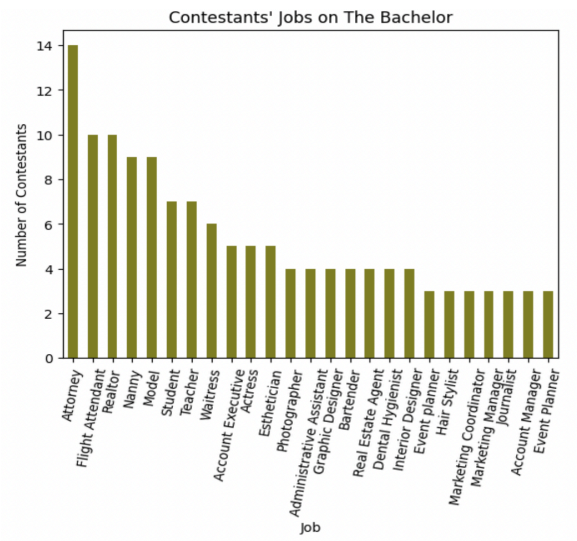
Visual Explorations:

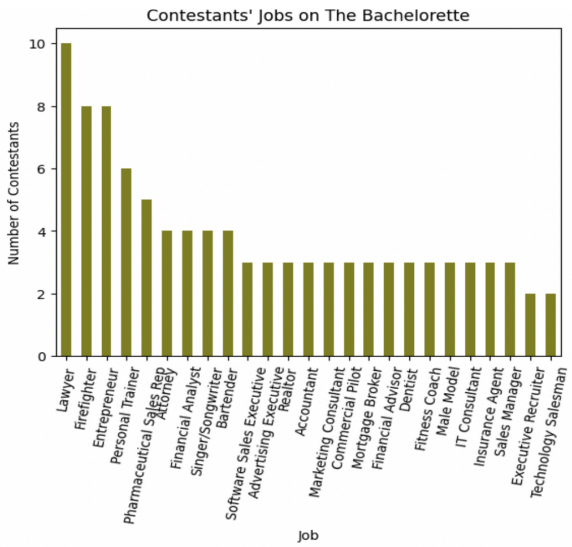
1. Distribution of contestants by show (left) & Distribution of what episode contestants are eliminated (right)





2. Distribution of jobs (Bachelor (Left) vs Bachelorette (Right))
    a. The occupations with the highest representation in the Bachelor were attorney, realtors, and flight attendants; in the Bachelorette, they were lawyers, fire fighters, and entrepreneurs.





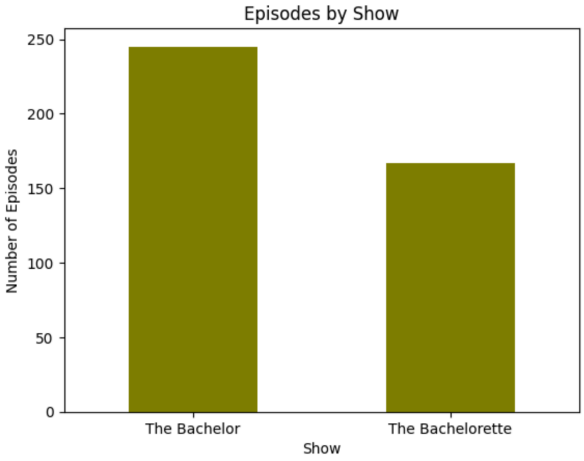
<u>Episodes EDA</u>

Descriptive Statistics:

1. Average viewers per episode: 8.61M
2. Average episode rating: 2.6 out of 5 stars
3. First episode air date represented: August 3, 2011

Visual Explorations:

1. Distribution of episodes for each show



Analytic Techniques

Exploratory Data Analysis: “Contestants” Dataset

Handling Null Values: We counted the number of NA values found in the contestants dataset, summing the totals using **.isna()** and **.sum()**. We decided to drop the rows with NA values in the dataset, which was a very small amount.

Categorical Variable Exploration: We first plotted the number of contestants against each show using matplotlib, showing that the bachelor had a higher total number of contestants. We used a **bar plot** to show this distribution and utilized the **value_counts()** function to add up the amount of contestants per show. Matplotlib functions such as **.plot()**, **.label()**, and .**show()** were used to create this bar plot visualization.

Variable Distribution: Matplotlib visualization techniques were used to plot the **.value_counts()** of the eliminations throughout each week, showing that week 1 had the greatest number of eliminations, following the format of the show. Overall, the distribution allowed us to understand how eliminations vary throughout time and episode number.

We looked into the contestants' jobs on *The Bachelor*, utilizing the "Show" and "Job" columns as well as the **.value_counts()** function to count the number of contestants who occupied different roles. Plotting these distributions allowed us to understand the difference in jobs between the two shows and what kinds of occupations were drawn to participating in the shows

We also added up the number of **unique** names, hometowns, and jobs throughout the two shows using the "Name", "hometown" and "Job" columns in the contestants dataset. This allowed us to conceptualize the data more efficiently.

Exploratory Data Analysis: "Episodes" Dataset

Null Value Handling: For this dataset, we first used **.isna()** and **.sum()** to find the number of missing values per column. We used **.drop()** to eliminate unnecessary columns and **.dropna()** on the rest of the dataset to get rid of these missing values.

Data Standardization and Preparation: We then standardized the "Rating" column to make sure that the data was consistent for manipulation. We removed quotes from the title column (using **.str.replace()** in **pandas**), and changed the "Air Date" to datetime (using pandas **pd.to_datetime**), alongside renamed some of the columns for readability using **.rename()** and **.replace()**. These functions allowed us to create cleaner, more understandable data.

Exploratory Q1 - Which version of the show has more viewers?

Descriptive Statistics: Calculated averages, medians, and standard deviations for key features such as age, **rating, and number of viewers**. For these calculations, we utilized Pandas function to **.groupby()** certain features and carried out the mathematical algorithms using python modules like **.mean(). Seaborn** was used to create visualizations of these relationships.

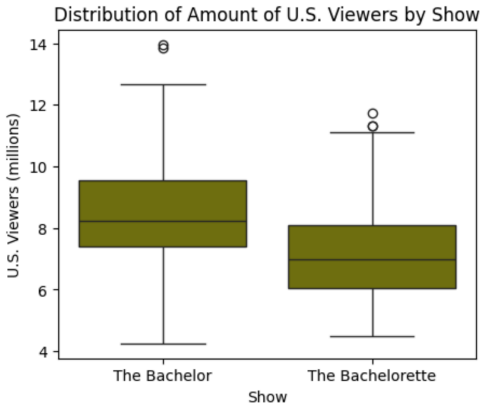
We found the distribution of US viewers per show using a box plot visualization in **seaborn**. This showed that *The Bachelorette* had a much smaller spread and less viewers than *The Bachelor* on average. These

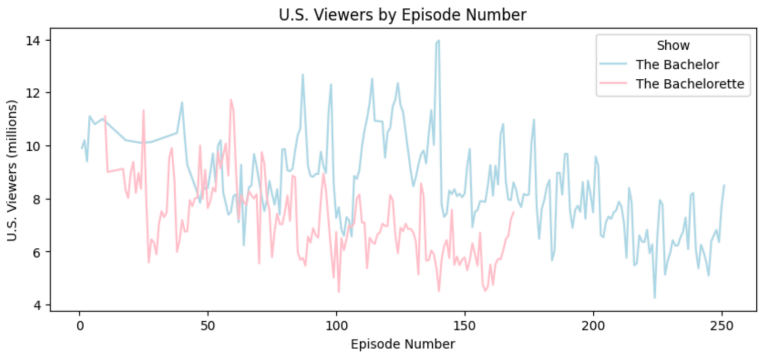
plots show significant differences between the average number of viewers for each show. This makes sense, as The Bachelor has been running for a longer time and has more dedicated viewers.

Correlation Analysis: Examined relationships between variables like **"Episode Number"** and **"U.S. Viewers"**. We used a **seaborn** line plot with the hue being set to show to label and create the graph. Overall, this distribution showed that as episodes increased, the number of viewers in the US decreased. The most viewed episodes are typically right before finales, when the Bachelor/Bachelorette makes their crucial decision of the season.

Visualizations:

- Boxplot of episode views by show (Left) & Line plot of views by episode (Right)





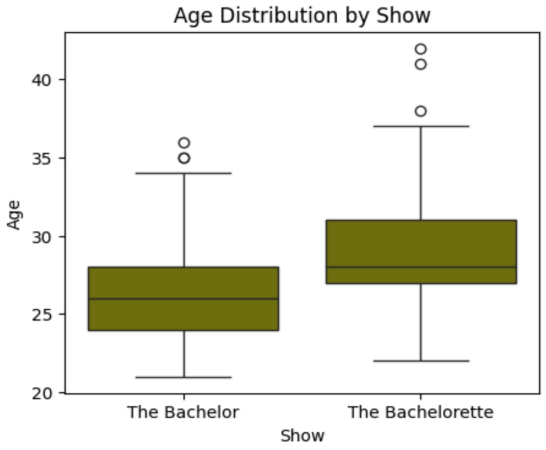
Exploratory Q2- Are ages of contestants similar between *The Bachelor* and *the Bachelorette*?

Distribution Plotting: We grouped the "Contestants" dataset by show to show the distribution of ages per show. We used a **seaborn** boxplot to show this distribution, finding that *The Bachelorette* had a wider spread of ages and more outlier ages than *The Bachelor*, as well as older overall. The Bachelorette contestants were typically between 27 and 31, while The Bachelor contestants were typically between 24 and 28.

Statistical T-Test Modeling: We performed a **t-test** on *The Bachelor* and *The Bachelorette* datasets to get a better idea of the statistical significance of the age differences found during our distribution plotting analysis. Using the **scipy.stats** module, we found the **T-statistic** and **P-Value** for the "Age" columns of the contestants on each show. This showed us that the difference in ages between the two shows was statistically significant.

Visualizations:

1. Distribution of ages by show



Exploratory Q3: Are there geographical patterns in contestants' hometowns and do these shift between shows?

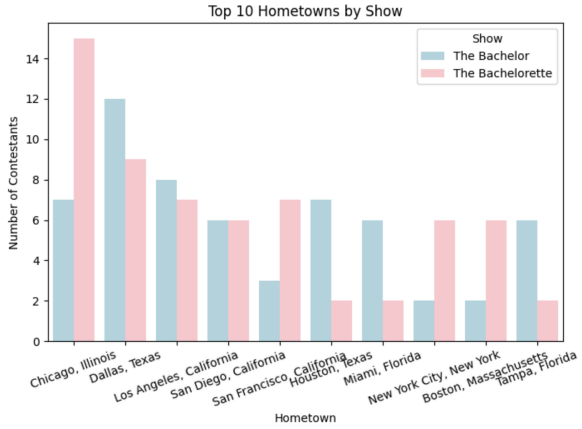
Data Merging: We **merged** the "Episodes" and "Contestants" datasets on the "Show" and "Season" columns. We created a "year" column to further analyze trends throughout the years on the two shows. We used **.drop_duplicates()** in our merged dataframe to prevent errors in our analysis. Through this merging technique, we were able to better manipulate the data for further analysis.

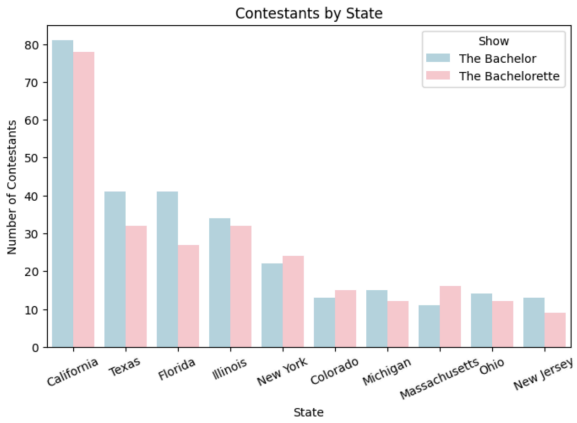
Visualization Analysis: We found the top 10 hometowns per show through manipulating the merged data frame and counting the number of contestants per hometown. We graphed our findings using a **seaborn bar chart** with **hues** set to the different shows, allowing us to see the popularity of different hometowns.

We repeated this process for the state values in each show. We used the same **seaborn countplot** functionality to plot this distribution, which showed the top three states to be Florida, Texas, and California, with California outnumbering the total number of contestants by a large margin. We also plotted the contestants numbers by state to better understand this distribution. We used **.value_counts()** to indicate the top ten states in our visualization. Some contestants were not from the U.S., but their hometowns were not common enough to affect the top states that contestants came from in either show. Chicago and Dallas were the most common states for a contestant to be from, but California was the most common state.

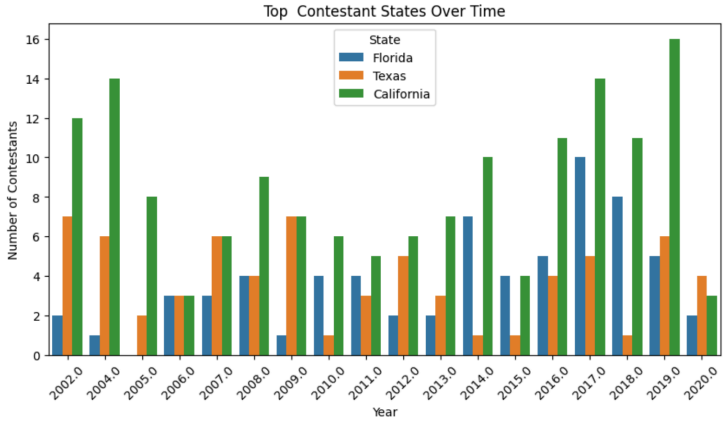
Visualizations:

1. Most common hometowns (Left) and home cities per show (Right)





2. Plot of how many contestants came from the top 3 states over time



Exploratory Q4: Are there other factors that are impacted by the Bachelor/Bachelorette?

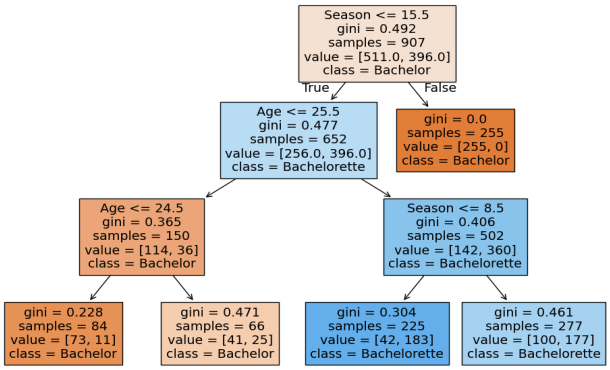
Logistic Regression: We used the "contestants" dataset to create a **logistic regression model.** We first prepped the data, separating the features (the age), from the corresponding label. We used the **logistic regression** model and **fit the x and y features** to the model. From this model, we found the summary, which showed the **p-value** and **coefficients** for the model, helping us to understand how age correlates with being a contestant on the show.

Decision Tree Model: To further analyze different factors pertaining to the shows, we used a Decision Tree model to consider features such as season and age. We first separated the data into **x and y variables**, using "Age" and "Season" in the x, and the "Is_Bachelorette" column for the Y. Using the **classifier** with a **max_depth of 3**, we fit the model and created a plot to show the splitting in the model.

We found that **Age** was the most important feature in determining whether a contestant was more likely to be on any given show, as the Decision Tree split on age. Contestants with younger ages were determined to be likely to be on *The Bachelor* whilst older ages (over 25.5 years old) were determined to be likely to join *The Bachelorette.*

Visualizations:

1. Plot of Decision Tree Modeling



Visualization Techniques

Bar Charts:

1. Distribution of contestants by occupation and location, alongside the number of contestants on each show using bar chart techniques.
2. Grouped the number of contestants by their elimination time, showing when contestants were more likely to be dropped during the show
3. Showed the average US viewers per show, alongside the number of episodes per show
4. Used to create a plot showing the Top 10 States by show winners

Box Plots:

1. Showed the distribution of viewers per show,
2. Used a boxplot to analyze the age distribution of each show

Line Graphs:

1. Depicted number of viewers through episodes in time
2. Found the number of contestants per season using a line plot, alongside the number of episodes and average ratings per season

Findings & Interpretations

Key Insights:

1. On average and overall, episodes of The Bachelor get more U.S. viewers than The Bachelorette
2. The most popular occupations included high-paying positions, such as Lawyer, Entrepreneur, Attorney, and Flight Attendant on both shows
3. The amount of viewers per episode decreased for both shows as episode number increased
4. Contestants for The Bachelorette tend to be older than contestants on The Bachelor with a statistically significant difference
5. Age seems to be an important factor in whether a past contestant was pulled from each show

6. The greatest amount of contestants come from California and Florida for both shows

Limitations & Challenges

Data Quality: When sorting through the datasets, we found there to be incomplete or inconsistent data from earlier seasons, which prevented us from being able to make particular comparisons. Some of the columns were misspelled, and many of the columns had null values, which we had to consider.

Overfitting: Due to the limited amount of factors presented in the dataset, there may have been overfitting within our models. There are many outliers, alongside irregularities that may have caused our models to not be trained in the most effective manner. Additionally, other factors, such as popularity on previous seasons and world events, would likely affect the data. Our models took a much more generalized approach and were trained using very specific variables, contributing to this overfitting. More data would need to be collected to get a holistic view of each show.

Generalizability: The insights found in our analysis are specific to the Bachelor/Bachelorette franchise and may not apply to other reality TV contexts. As a result, it is difficult to predict the impact of our work and its applicability to any other reality show.

Recommendations

For Networks: It is important for networks to focus on casting a wider net of ages per show to increase variety. Diversification can help with engagement, alongside showing interest over time. As shown by our analysis, one of the key issues is that particular job occupations, alongside age ranges and hometowns, were found more frequently in the two datasets. Older ranges were more likely to be cast on *The Bachelorette* and most contestants came from California and Chicago. *The Bachelor* had a more evenly distributed representation of hometowns than *The Bachelorette,* which may have contributed to the increased number of viewers for the show. Overall, by increasing diversity on the show of ages, hometowns, and demographics, networks may be able to appeal to wider audiences to help increase viewership.

For Contestants: We recommend contestants develop a strong social media presence before joining the show to emphasize their personal stories and help them stand out from others in the early seasons. Developing an online presence may increase the viewership of contestants over time. As shown in our analysis, views and ratings drop as seasons progress. If contestants are already influential, the show may maintain more consistent popularity over time.

Future Work: For a more refined approach, we would recommend conducting a sentiment analysis from social media accounts to understand audience insights. Machine learning language processing models could be utilized on social media datasets from contestants to measure engagement over time and how social media affects ratings and interest in the show. We also believe analyzing the differences in audience data across international versions of the show could be useful, as it would allow for a broader application of key findings, which would be more transferable to other shows and franchises.