The Bachelor
Exploratory Data Analysis

Group 5
Matthew Bailey
Alexandra Flores
Will Hughes
Vanessa Martinez

What is *The Bachelor*?

The
BACHELOR

Introduction

Our exploratory data analysis project is on the popular reality TV show *The Bachelor*.

What is *The Bachelor*?

- *The Bachelor* is an American dating and relationship reality television series that debuted on March 25, 2002, on ABC.
- The series revolves around a single bachelor who begins with a pool of romantic interests from whom he is expected to select a wife.
- During the course of the season, the bachelor eliminates candidates each week, eventually culminating in a marriage proposal to his final selection.
- On each Bachelor episode, the bachelor interacts with the women and presents a rose to each woman he wishes to remain on the show. Those who don’t receive a rose are eliminated.

https://en.wikipedia.org/wiki/The_Bachelor_(American_TV_series)

Introduction

Why analyze *The Bachelor*?

- We chose *The Bachelor* to gain insight about contestant demographics and to see if we can successfully make predictions regarding a contestant's success on the show.

- We focused on previous contestants' age, amount of time they were on the show, occupation, and where they were from, to conduct our analysis.

Data Information

Data Information

- We used datasets provided from Kaggle and Dataworld.
- 3 CSV files including information on contestants that we merged and narrowed down to the following:
  - Contestant Name
  - Age
  - Occupation
  - Hometown
  - Seasons of Show
  - Elimination Week of Contestant
  - Place (Rank)
- We used Openweathermap and Google Cloud/Maps APIs to collect location information

https://www.kaggle.com/datasets/brianbgonz/the-bachelorette-contestants
https://www.kaggle.com/datasets/rachelleperez/the-bachelor-vs-the-bachelorette?select=contestants.csv
https://data.world/amandanovak/bachelor-contestants-with-instagram-follower-count
https://openweathermap.org/
https://developers.google.com/maps

Data Information

- Unused information was Instagram followers, height, and male contestants

- We grouped occupation information into 18 categories:
  - Beautician, Education, Entertainment, Entrepreneur, Fashion/Design, Fitness, Finance, Health Related Field, Law Related Field, Media, Real Estate, Sales, Service Industry, Student, Transportation, Unemployed, and Veteran

- We created grouped bins for
  - Seasons
    - 1-7, 8-14, 15-21
  - Ages
    - 20-25, 26-29, 30-35, 36-39
  - Elimination Weeks
    - 1-4, 5-7, 8-10

Research Questions

Research Questions

- What was the average elimination week per age range?
- Has the age range and individual age of contestants changed over time?
- Do specific occupations lead to a contestant making it longer into the courtship?
- Do contestants from specific regions have a better chance of success?

Data Cleaning and Eliminations

Data Cleaning

- Imported all the data to Jupyter Notebook
- Two big merges resulting in a lot of duplicate columns
- Much like weeks one and two of the bachelor I got rid of all of the unnecessary information
- I filled in some N/a values in the ElimWeek Column and made some Age Bins

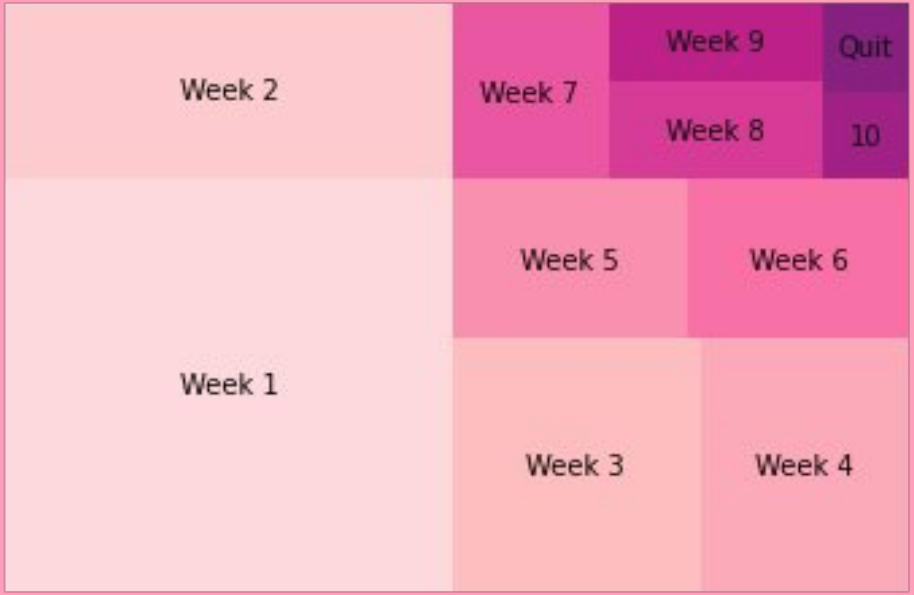
Elimination Breakdown

49.4% of all Eliminations happen in the first two weeks.

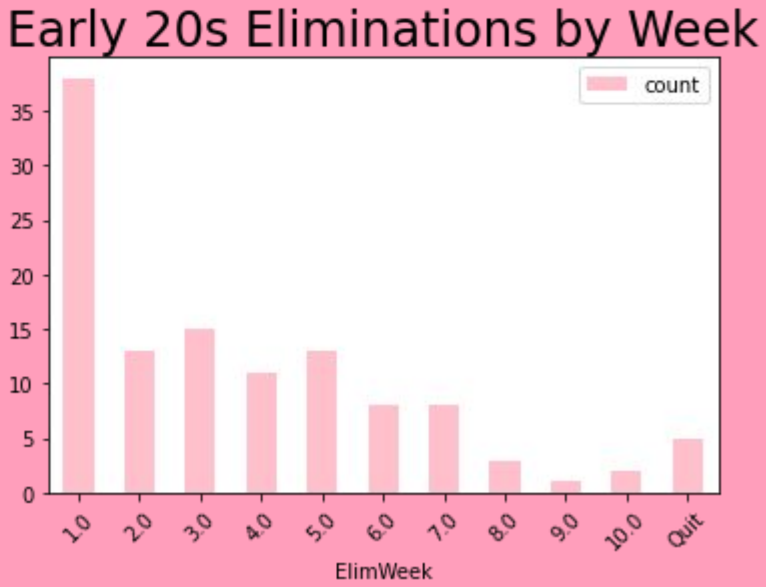
35% of all Eliminations happen in week 1 alone.

Only 4% of contestants have quit the show.

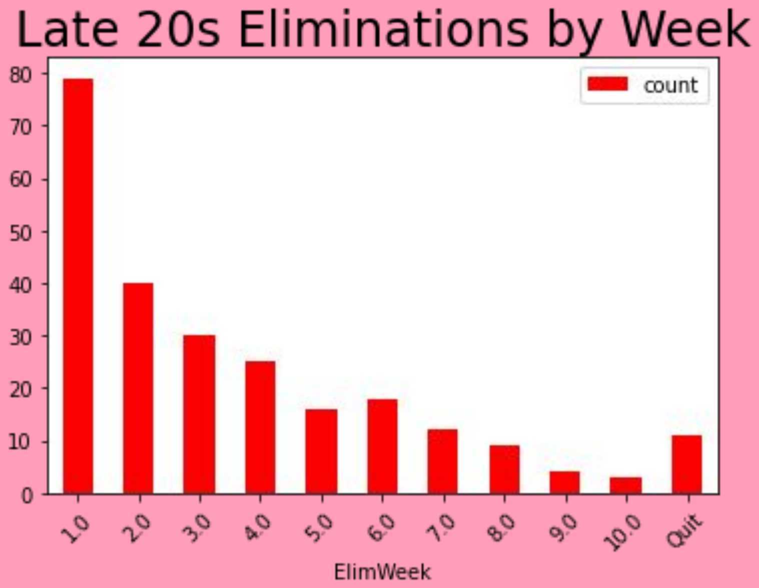
On average a given contestant has about a 5% or (1/20) chance of receiving the final rose



Eliminations in the 20’s Group

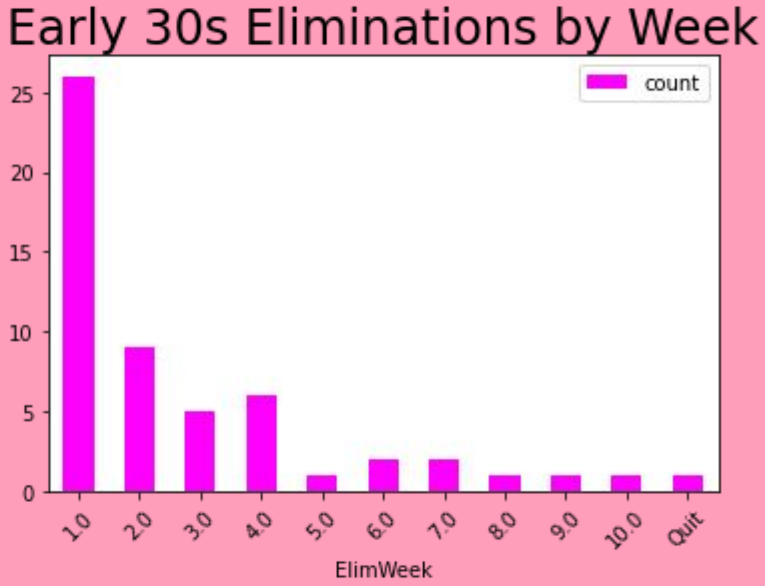


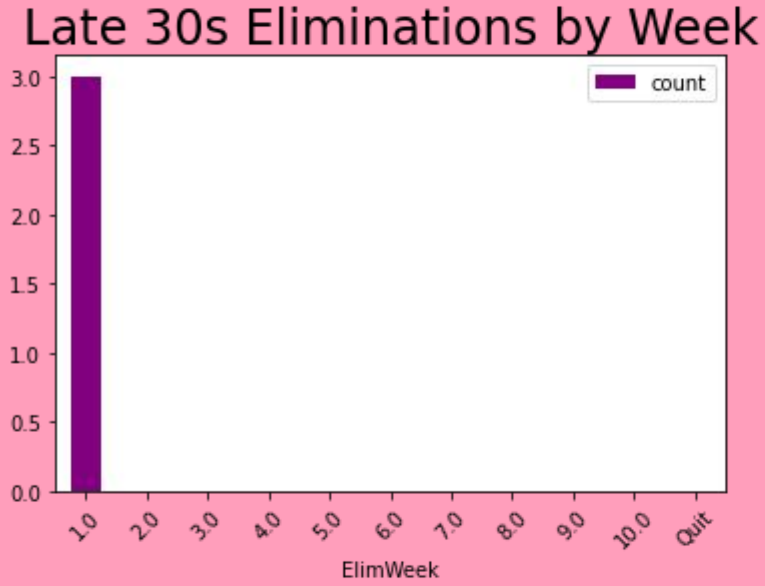
Early 20s Eliminations by Week



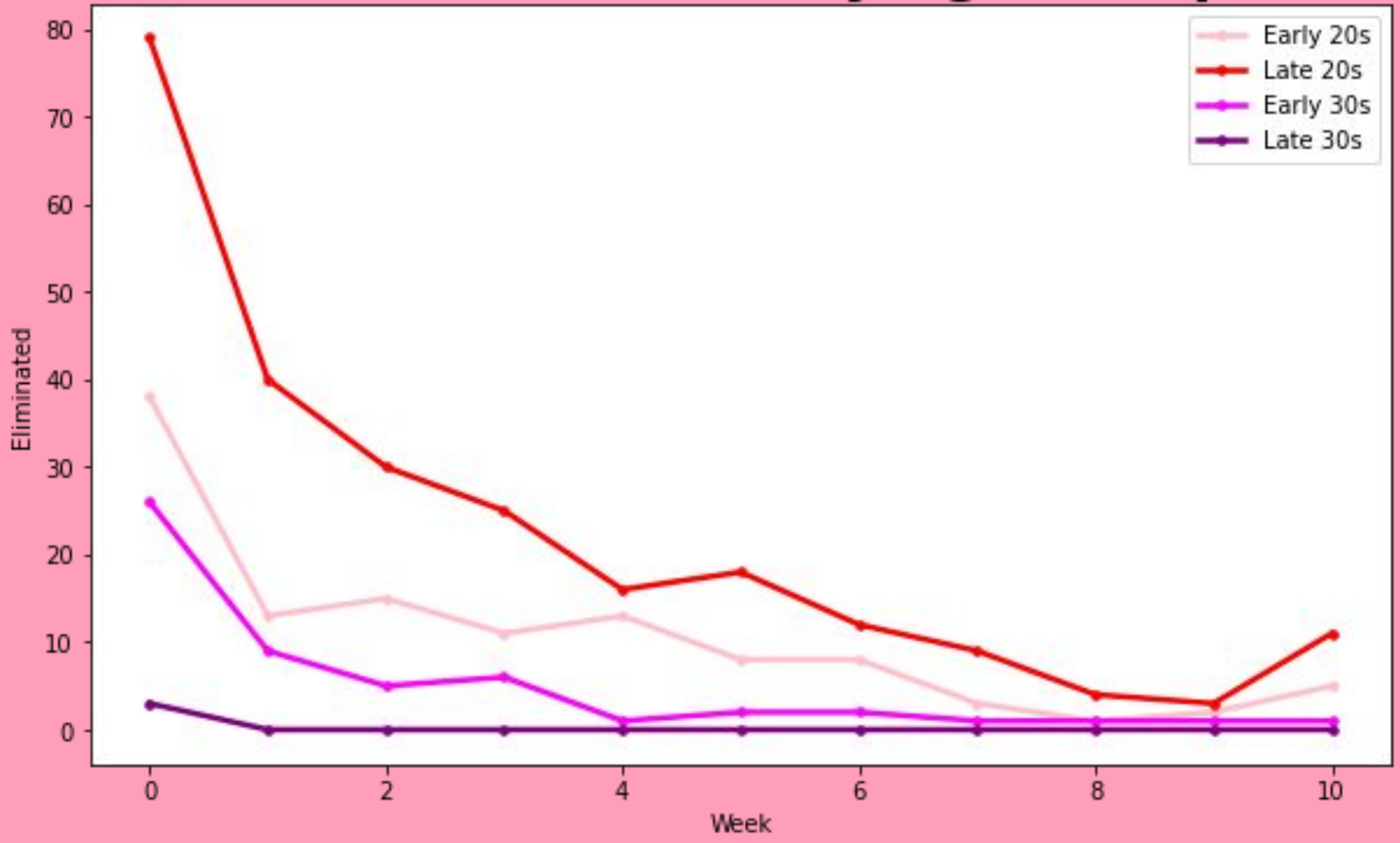
Late 20s Eliminations by Week

Eliminations in the 30’s Group


Early 30s Eliminations by Week


Late 30s Eliminations by Week

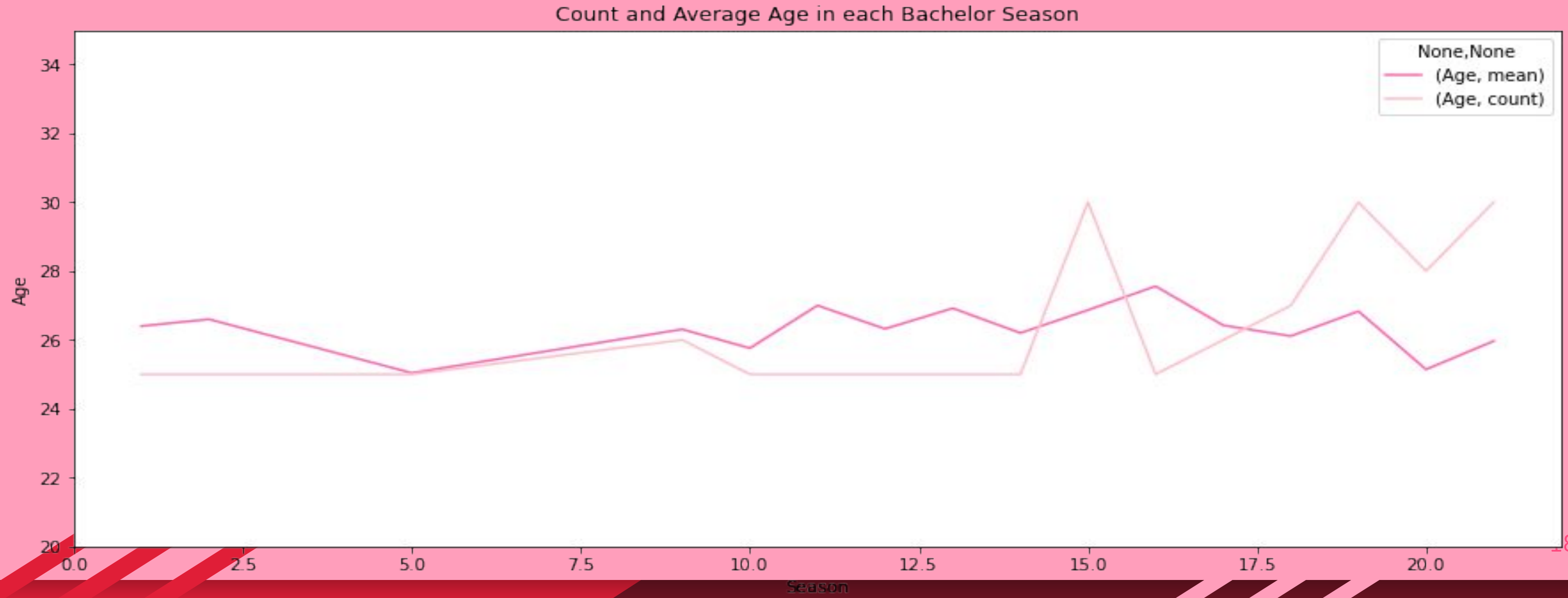
Elimination Week by Age Group



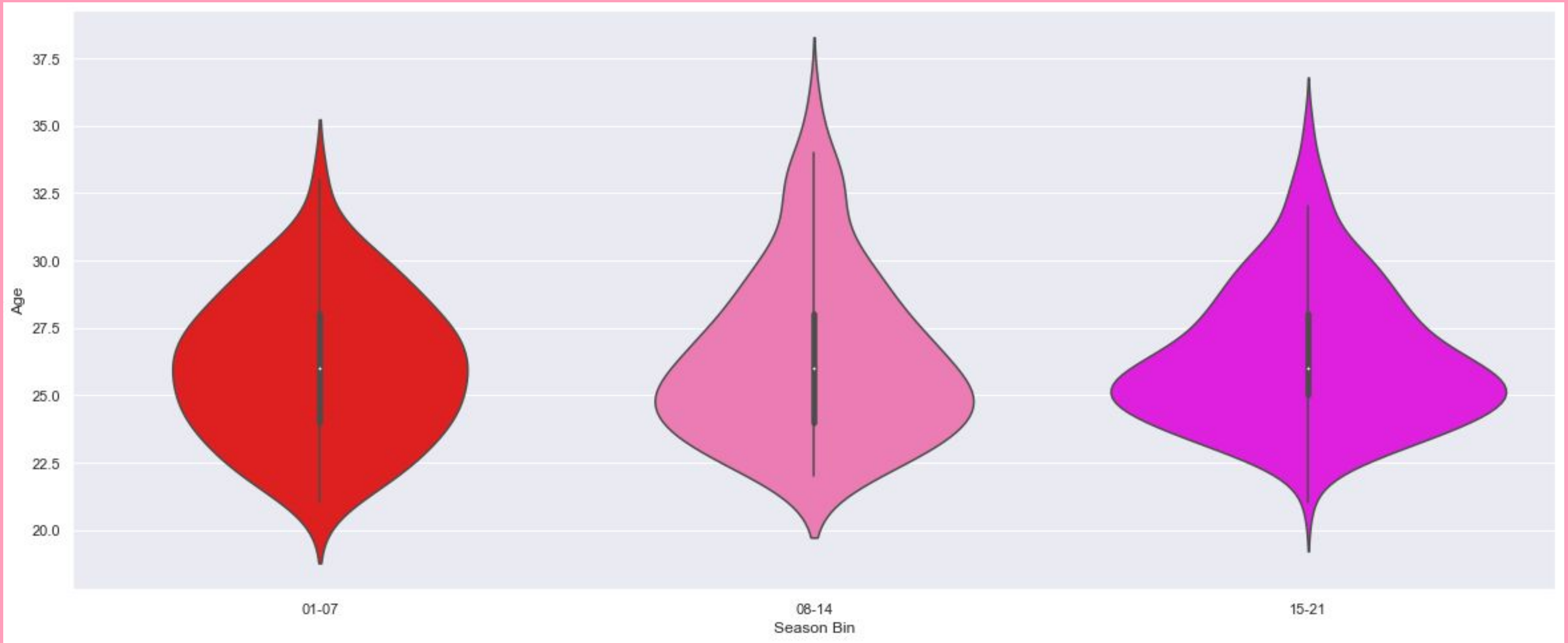
Age Analysis

Average Age & Count throughout the Seasons

Research Question: Did the average individual age trend up or trend down throughout the show?

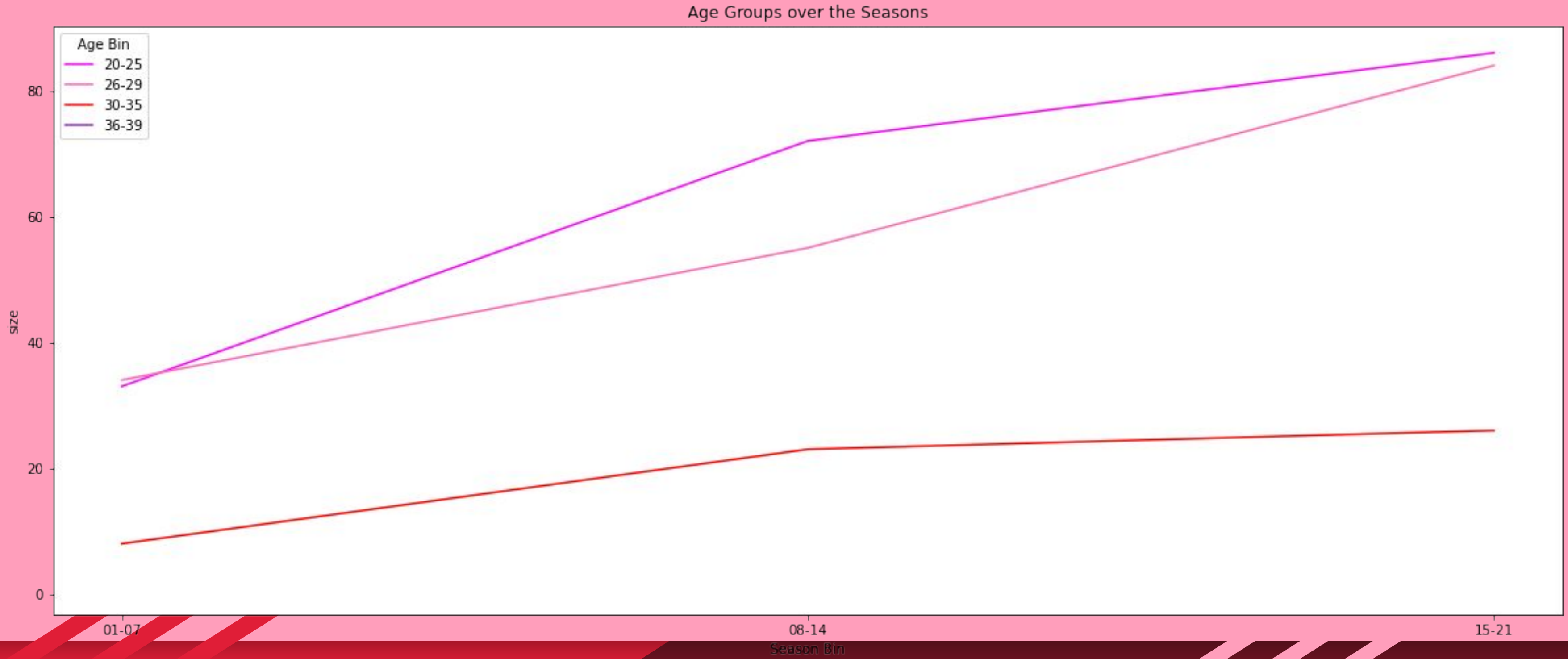


Age Variance

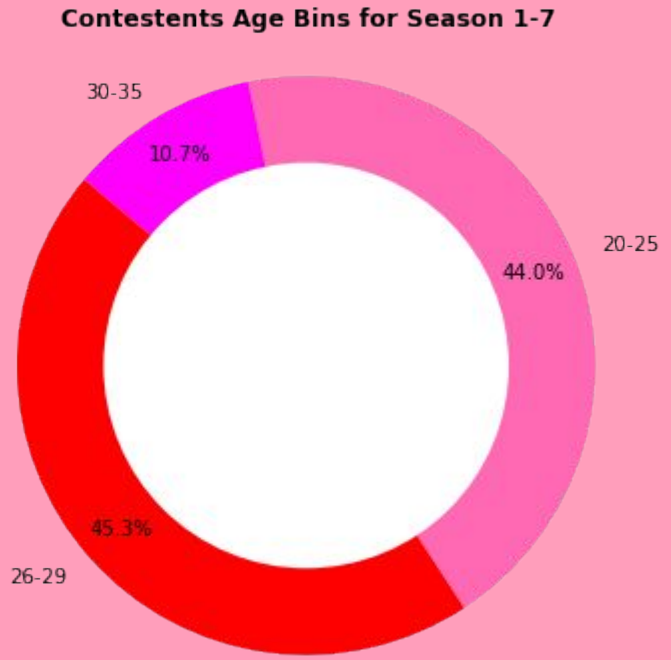


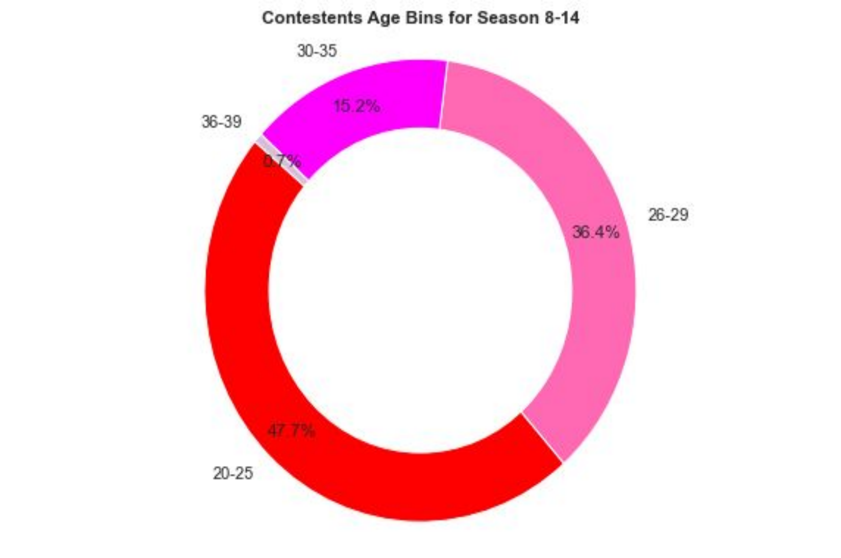
Age Groups

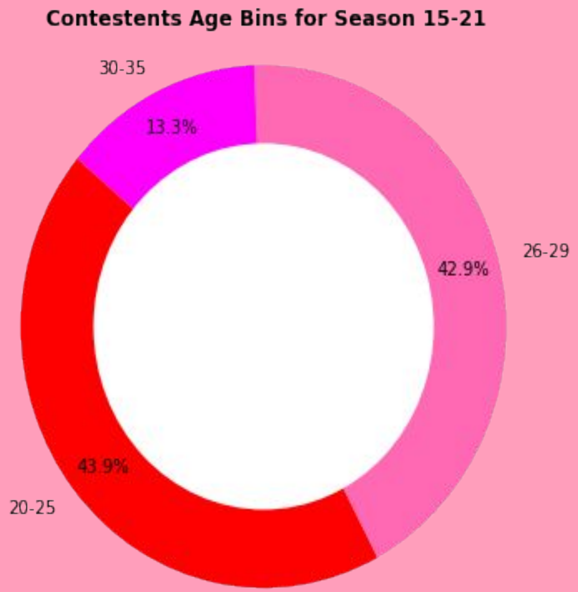
Research Question: Has the age range changed over time?



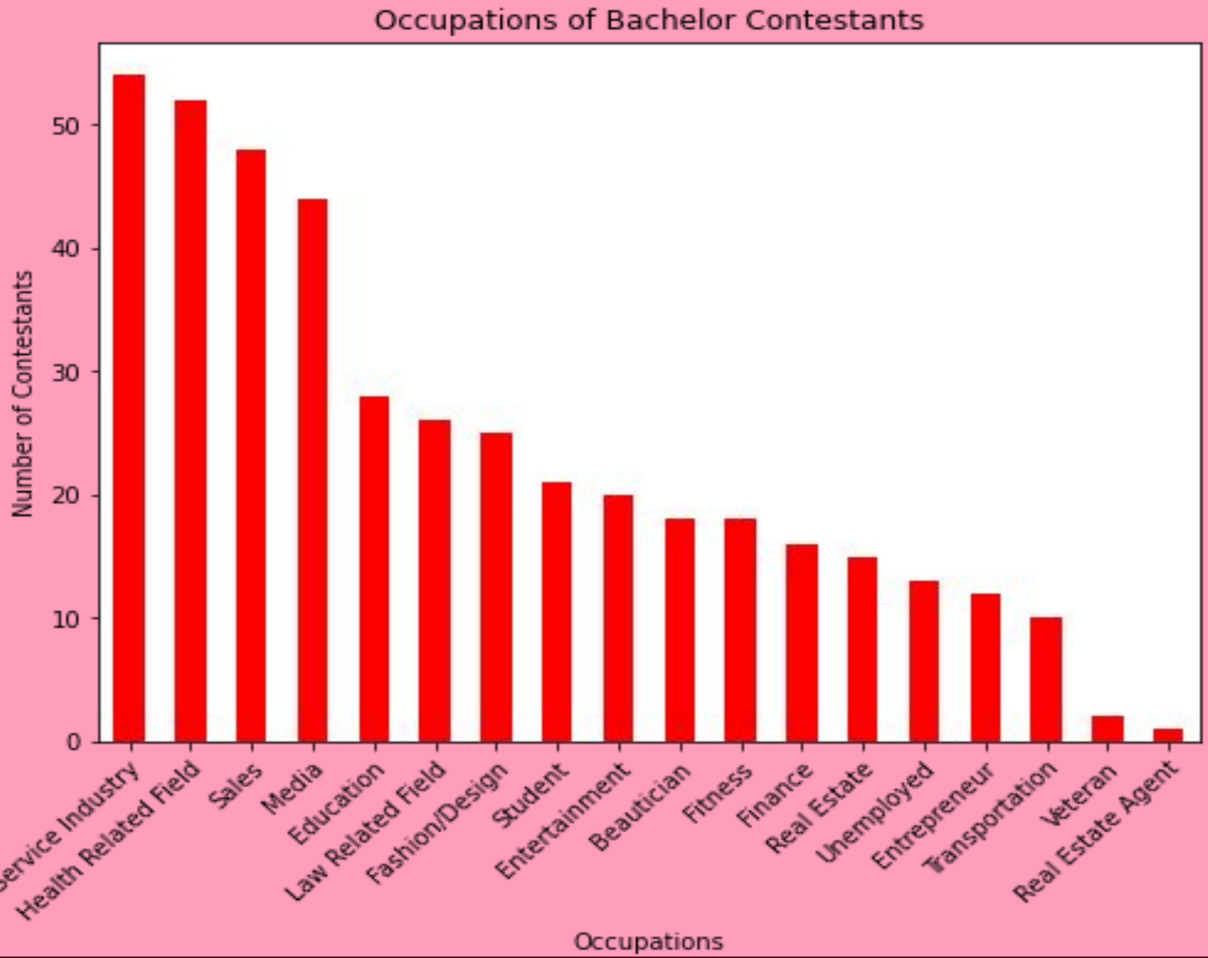
Age Groups Percentages



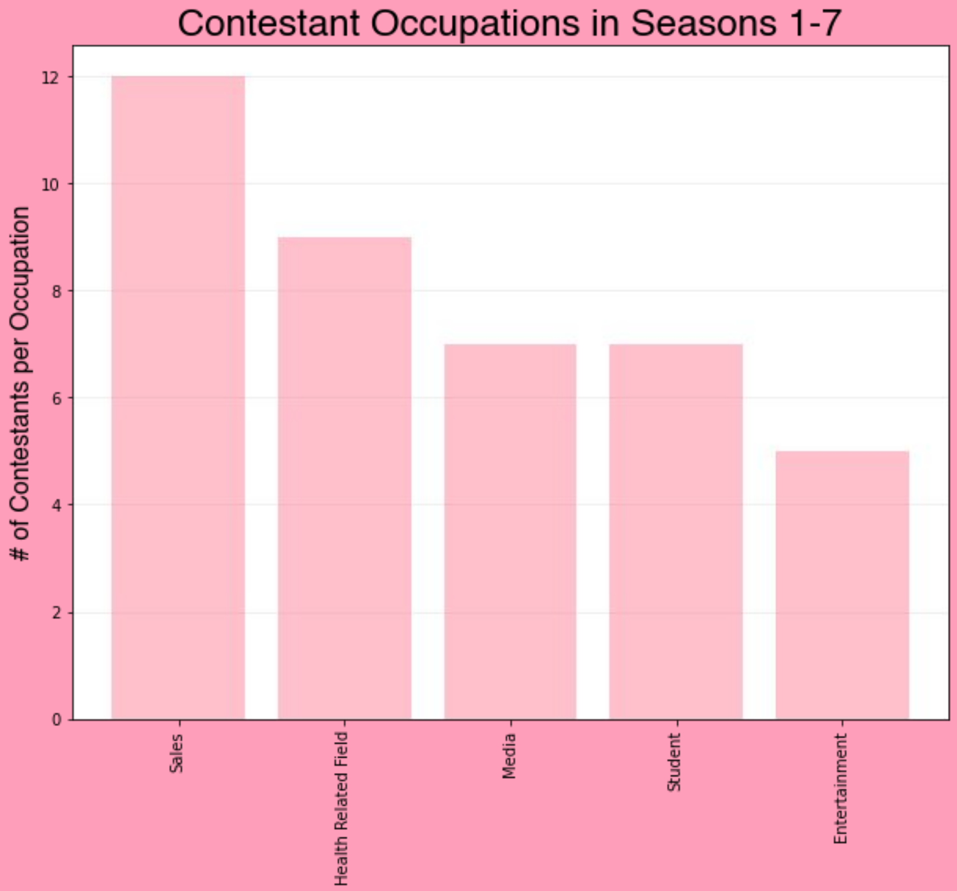

Contestents Age Bins for Season 8-14
30-35
15.2%
36-39
0.7%
26-29
36.4%
47.7%
20-25

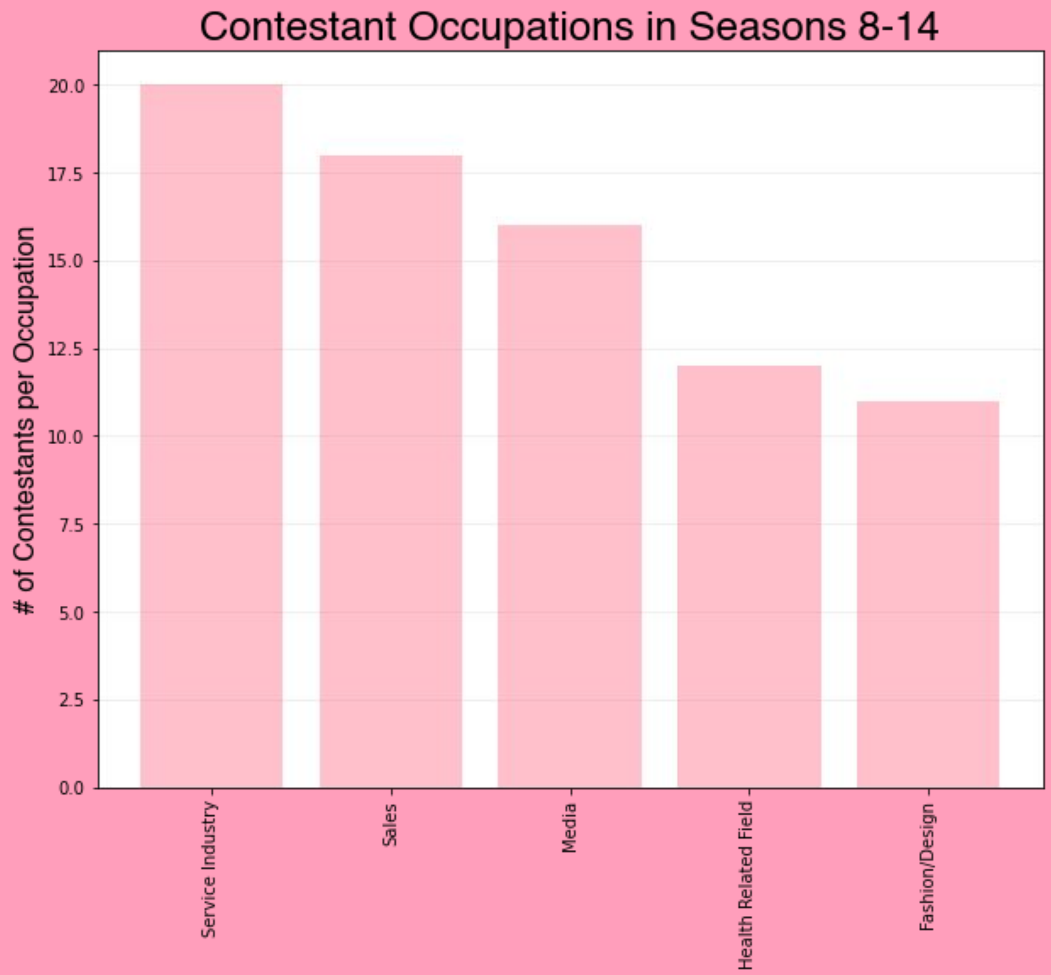

Contestents Age Bins for Season 15-21
30-35
13.3%
26-29
42.9%
43.9%
20-25

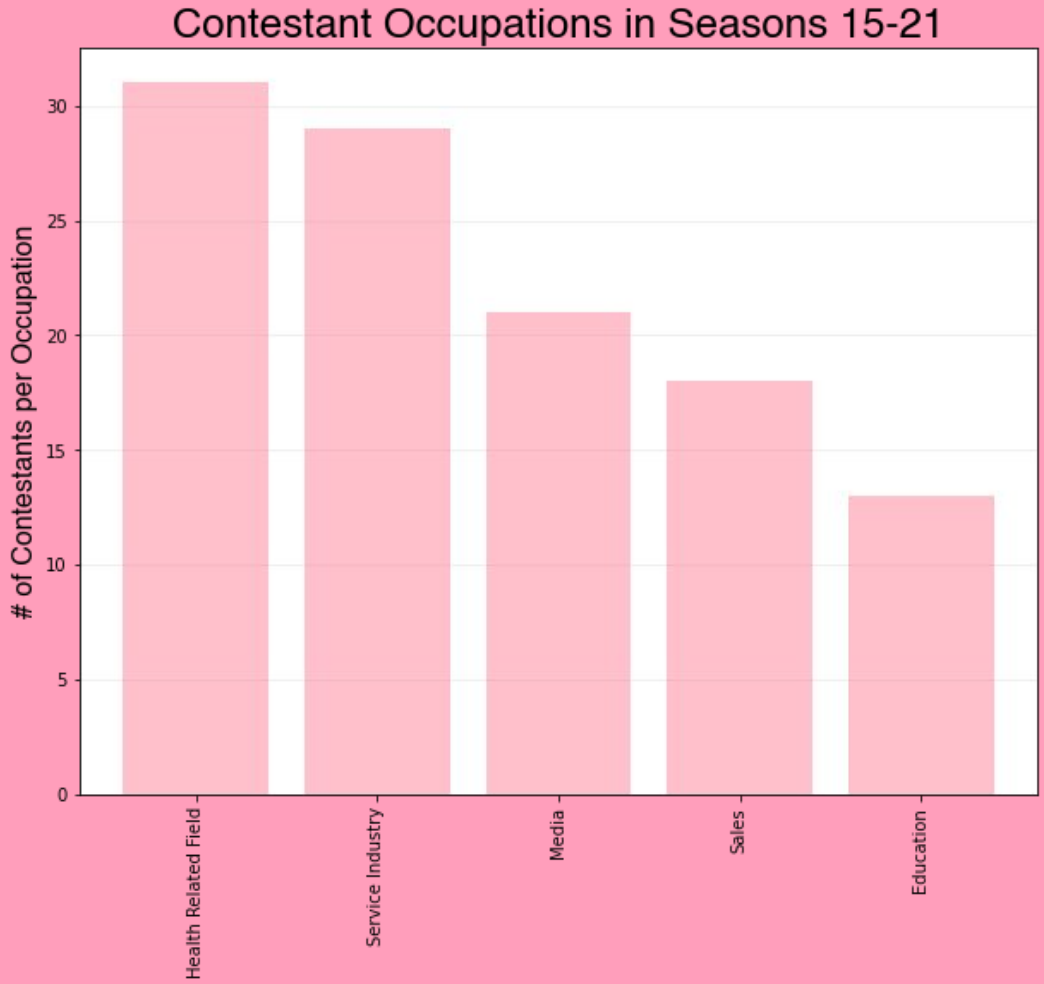
Occupation Analysis


Occupations of Bachelor Contestants
Number of Contestants
0
10
20
30
40
50
Occupations
Service Industry
Health Related Field
Sales
Media
Education
Law Related Field
Fashion/Design
Student
Entertainment
Beautician
Fitness
Finance
Real Estate
Unemployed
Entrepreneur
Transportation
Veteran
Real Estate Agent

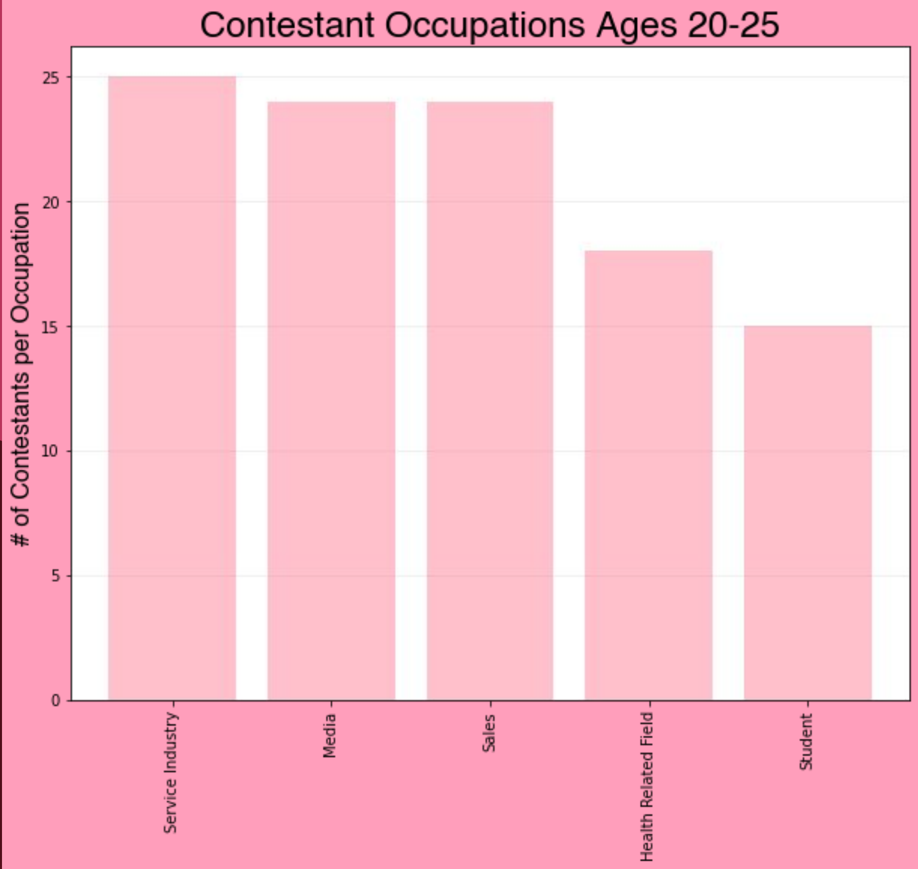
Most Popular Occupations in Earlier Seasons

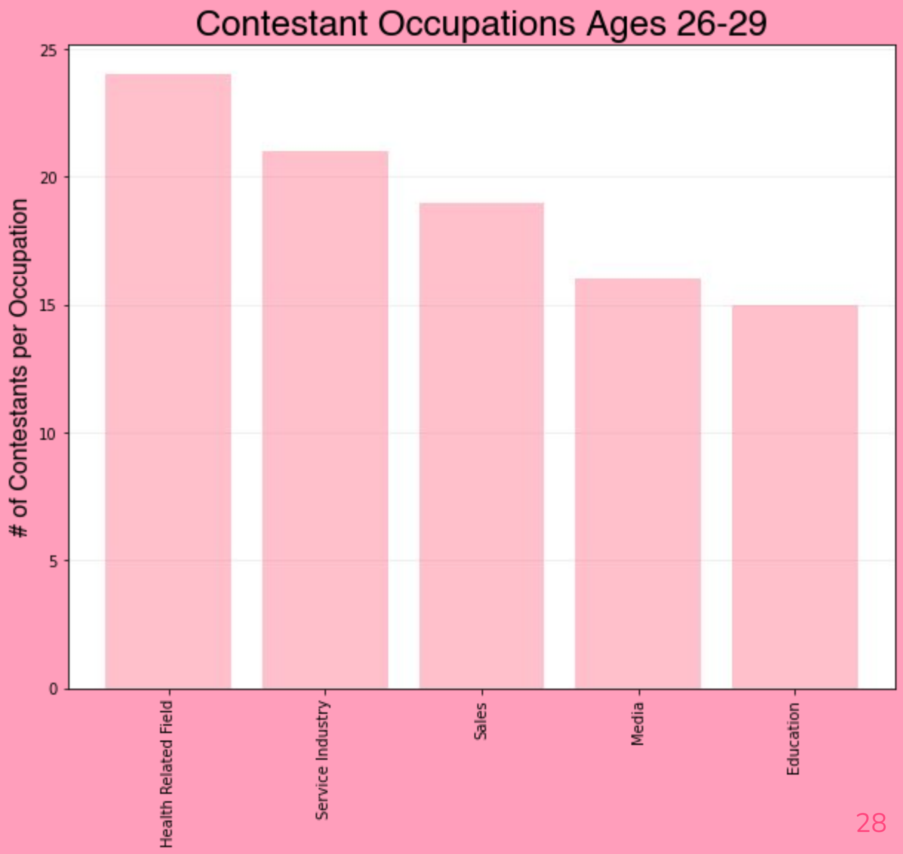



Contestant Occupations in Seasons 8-14
# of Contestants per Occupation
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Service Industry
Sales
Media
Health Related Field
Fashion/Design

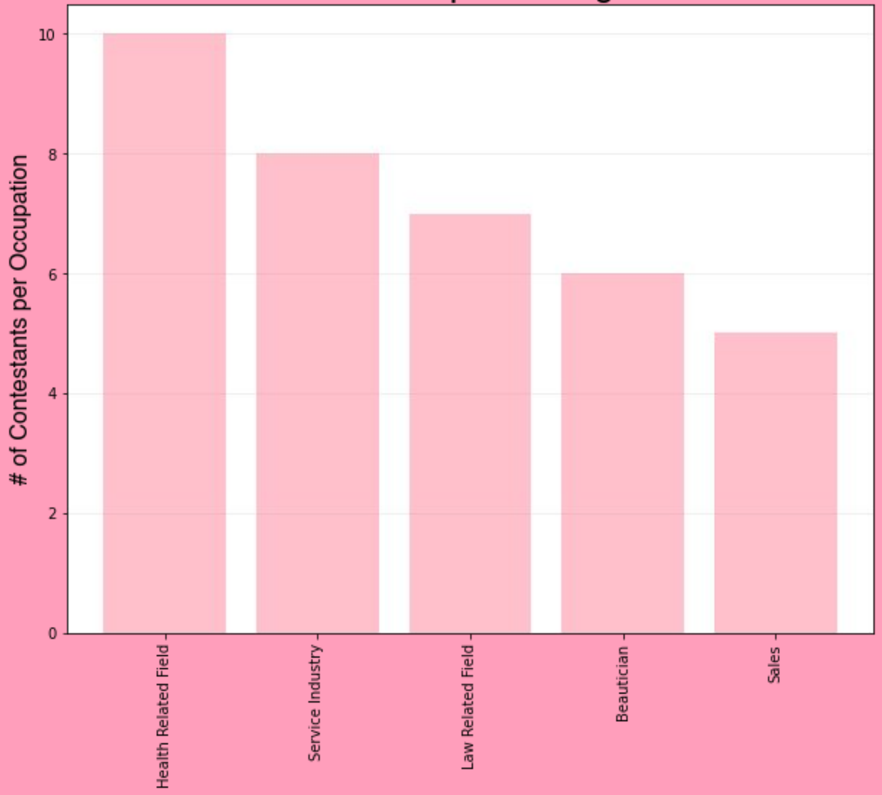

Contestant Occupations in Seasons 15-21
# of Contestants per Occupation
0
5
10
15
20
25
30
Health Related Field
Service Industry
Media
Sales
Education

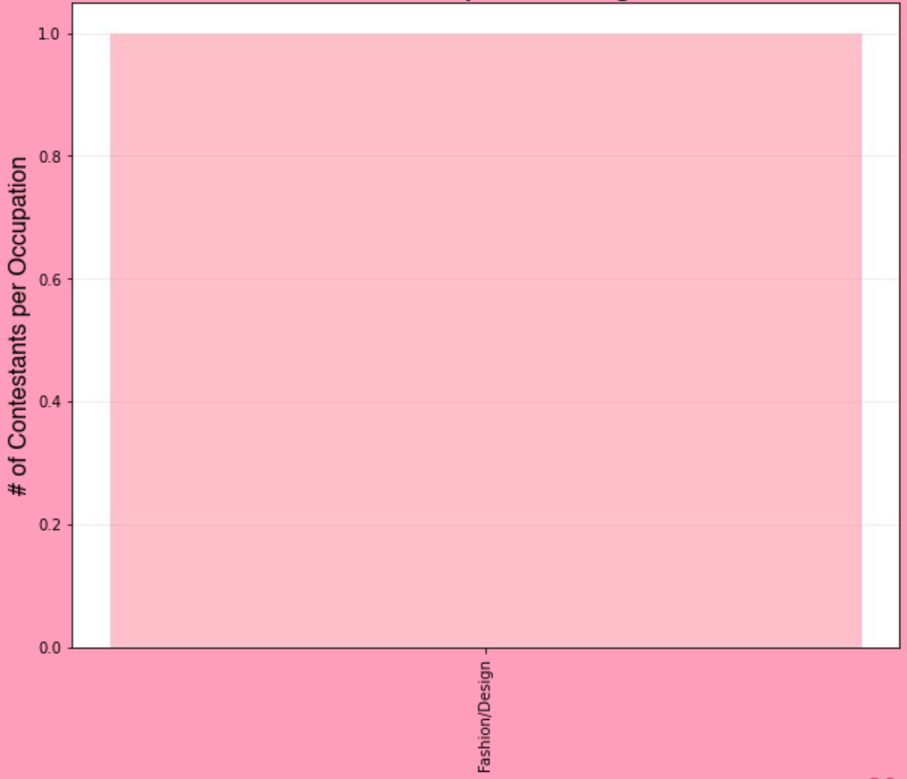
Most Common Occupations Based on Age



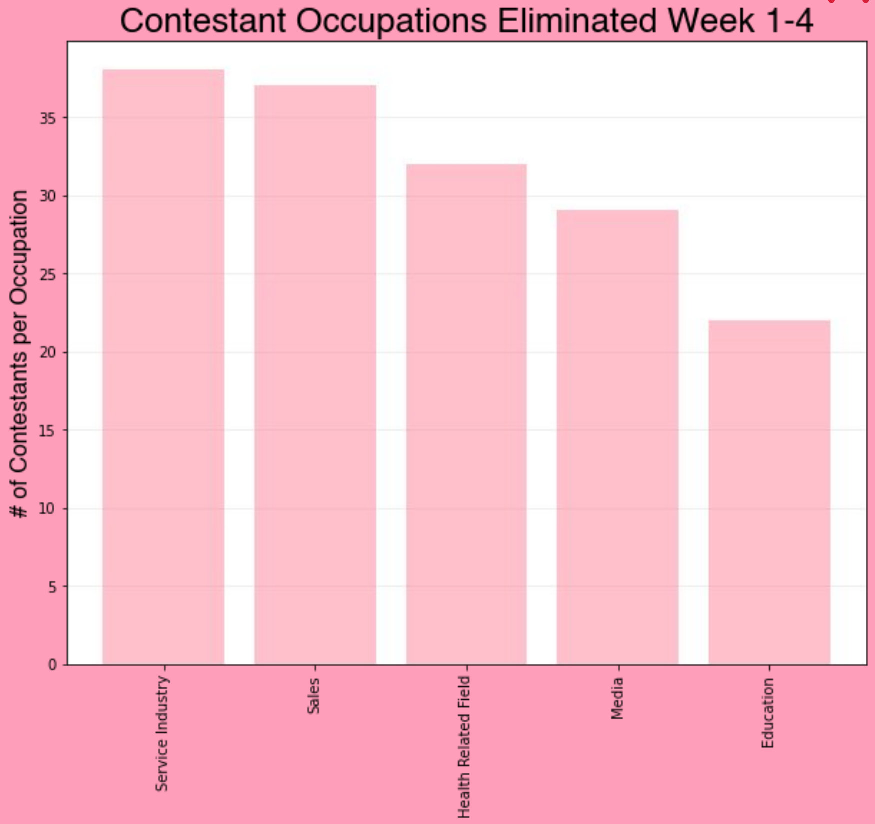


Most Common Occupations Based on Age

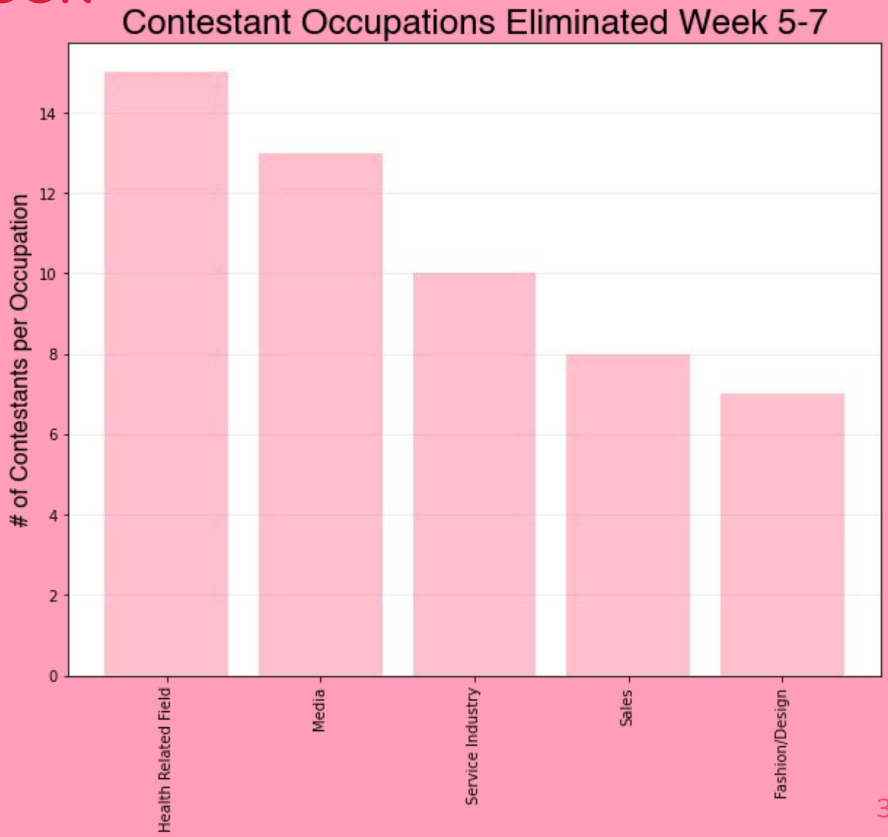

Contestant Occupations Ages 30-35


Contestant Occupations Ages 36-39

Most Common Occupations Based on Elimination Week

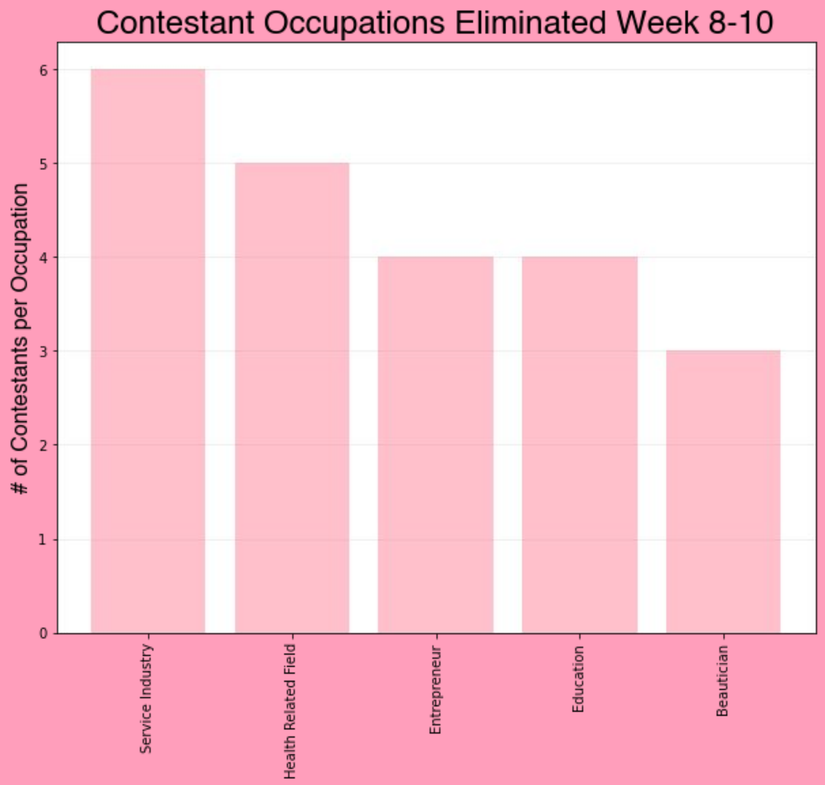


Contestant Occupations Eliminated Week 1-4



Contestant Occupations Eliminated Week 5-7

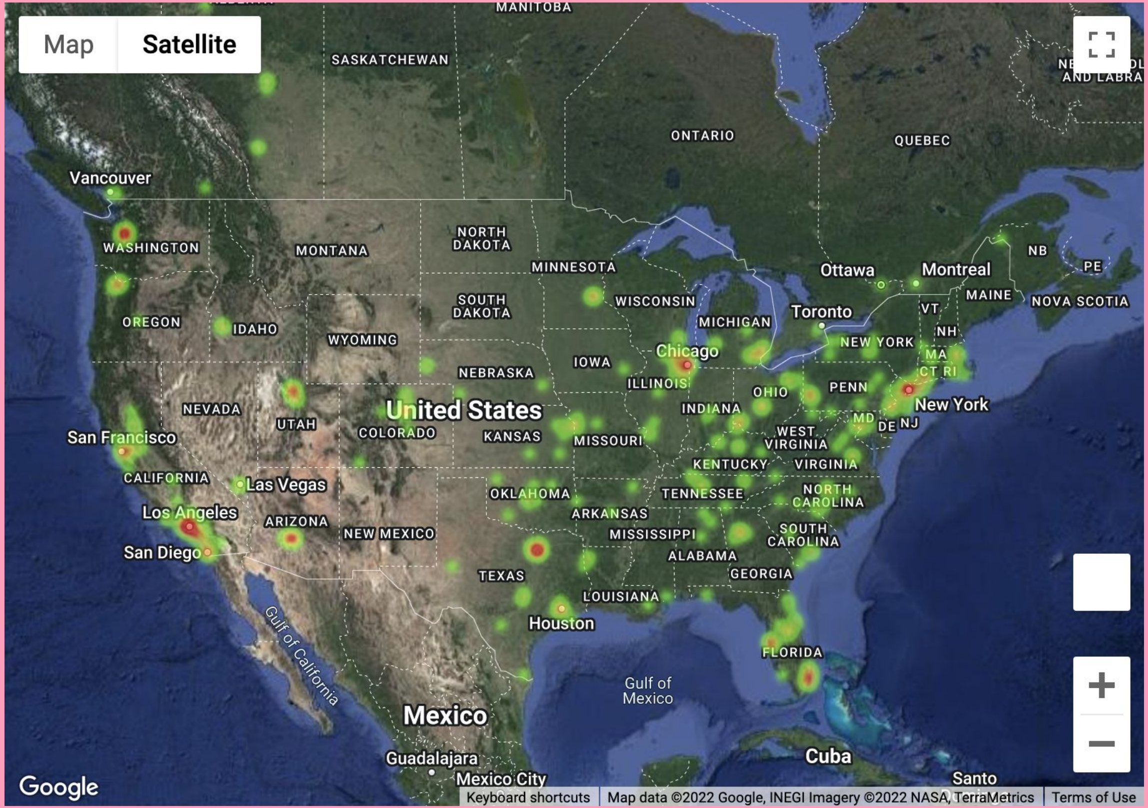
Most Common Occupations Based on Elimination Week

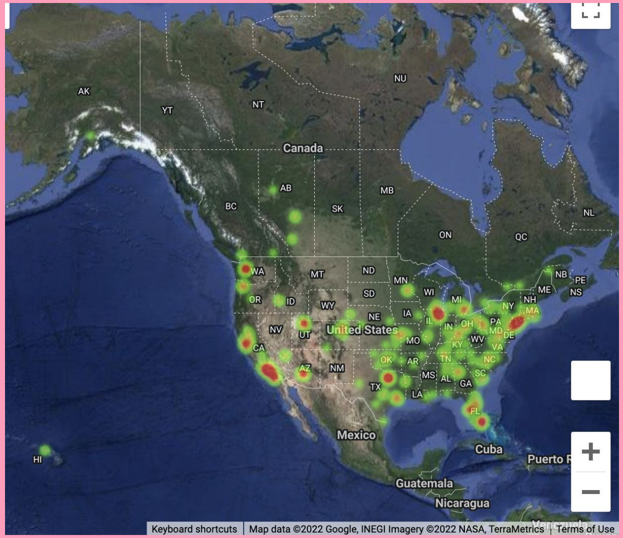


The most common occupations stayed consistent throughout the elimination process. Service Industry, Education, and Health related occupations were repeatedly seen throughout the season. Therefore occupation does not affect how far you would make it in the Bachelor courtship.

Location Analysis

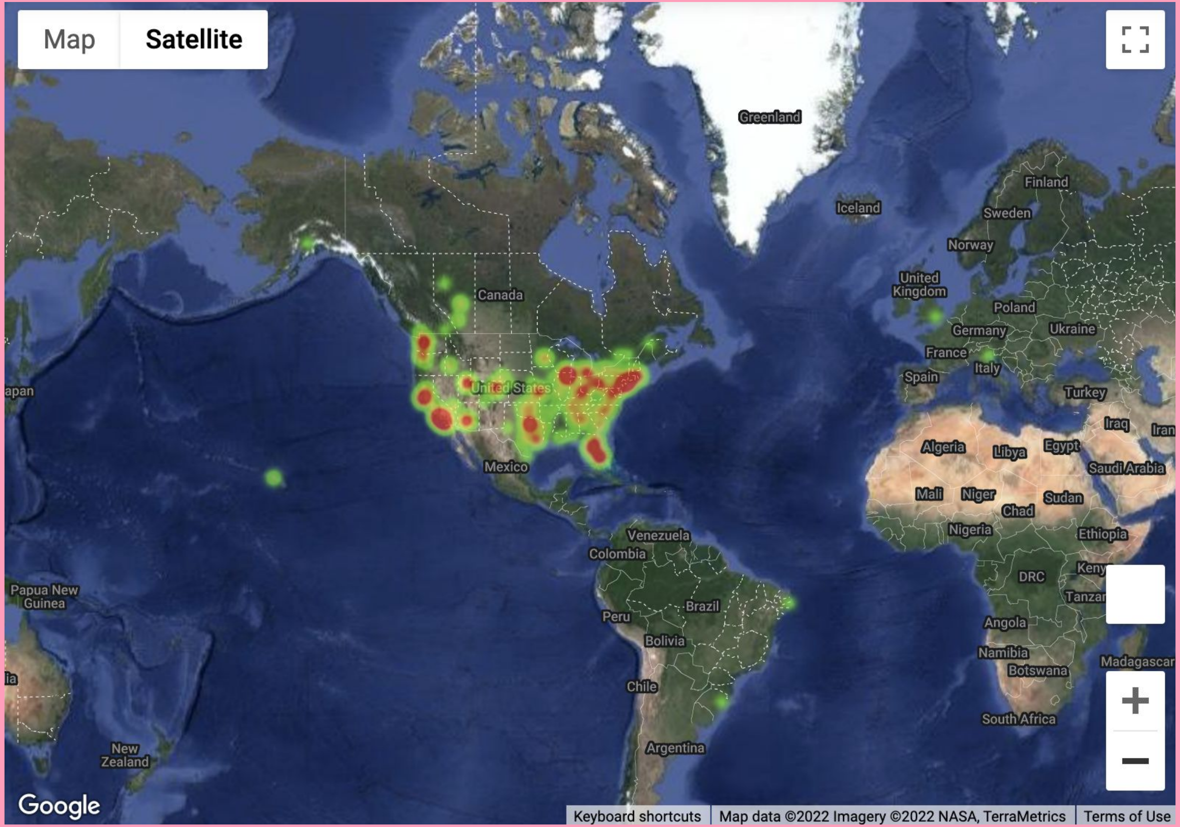
Locations Of All Contestants



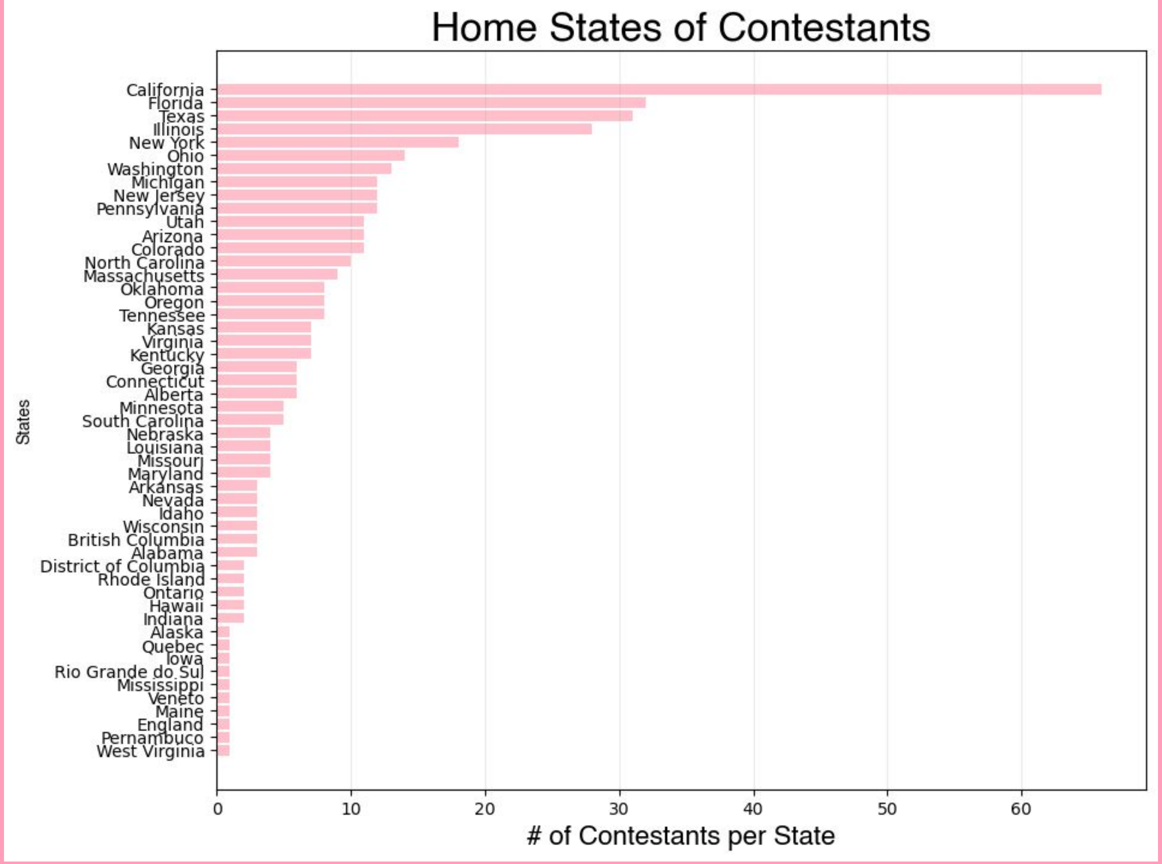


Majority are from the United States

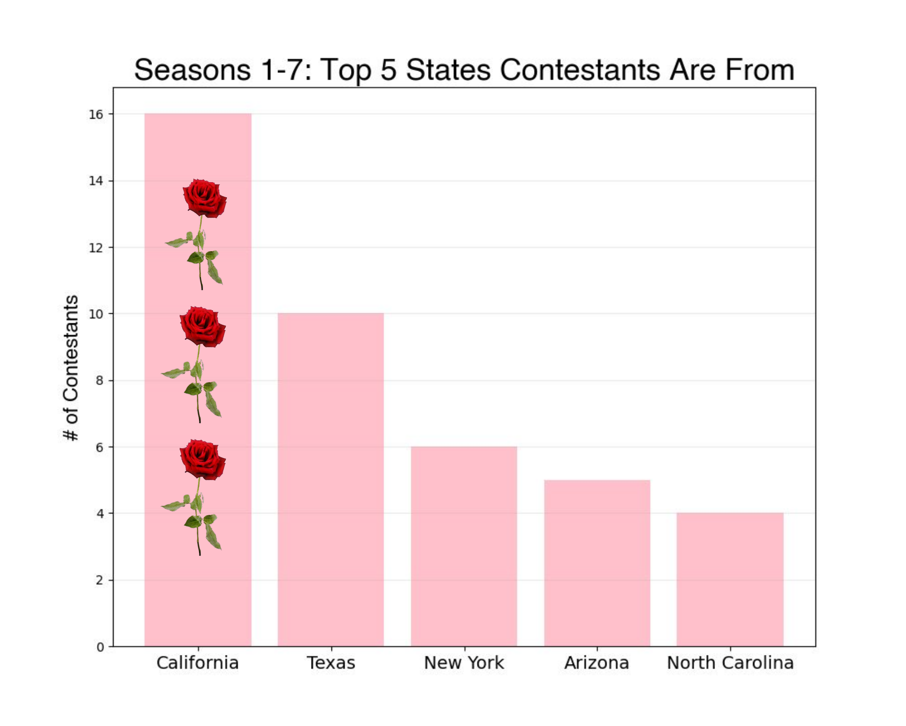
Locations Of All Contestants

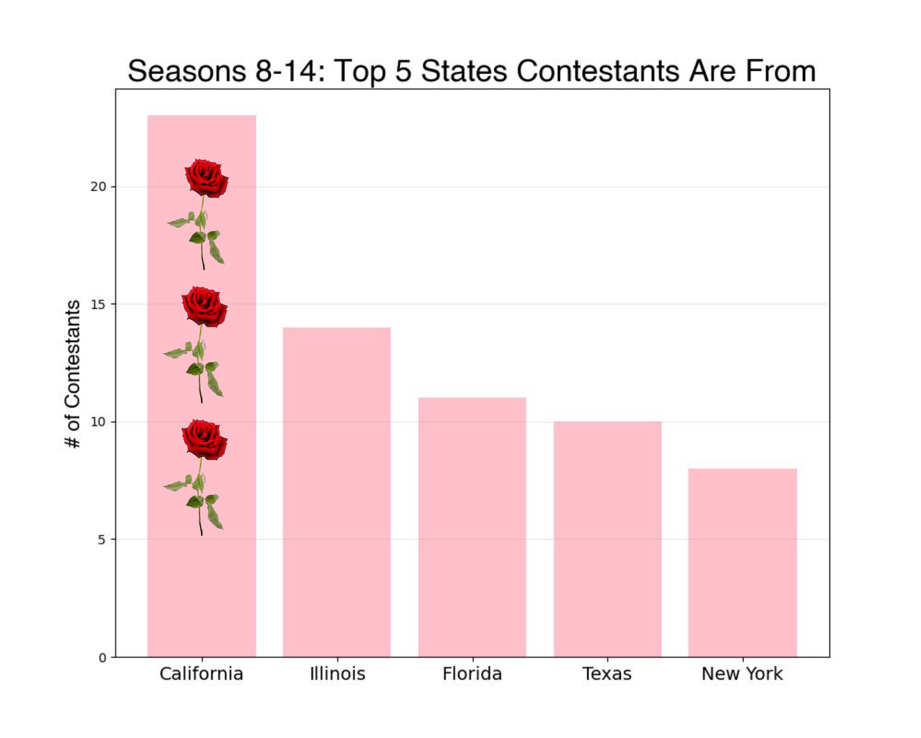


Locations Of All Contestants

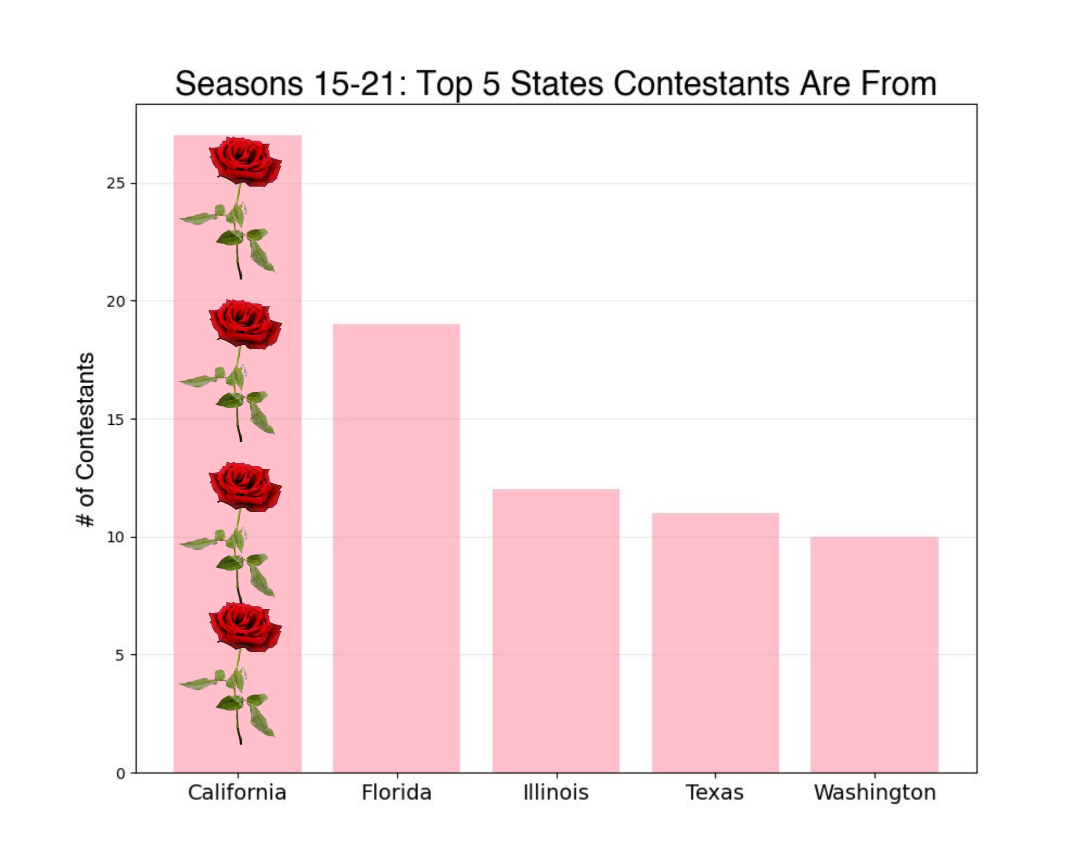


Top 5 Contestant Locations By Seasons

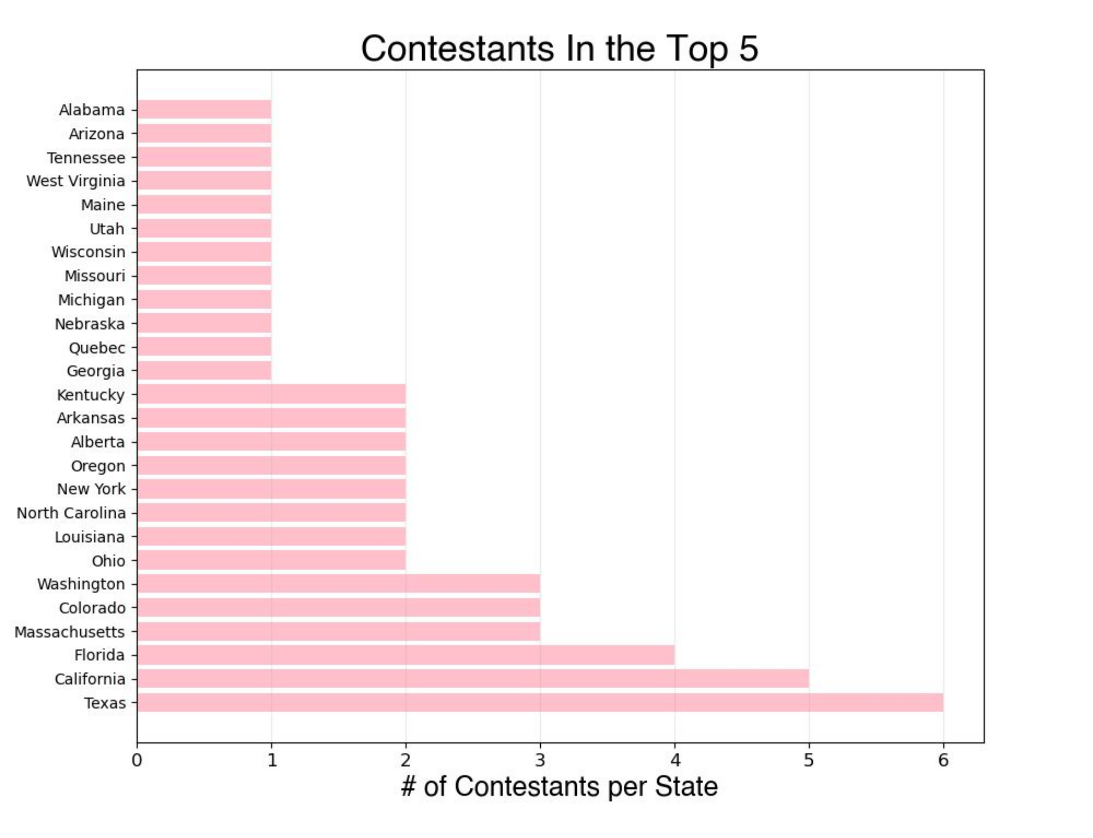


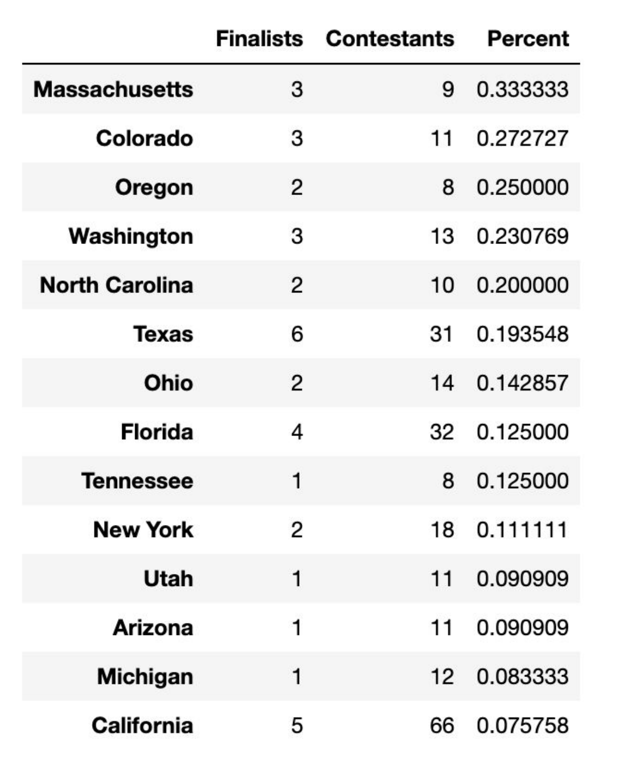


Top 5 Contestant Locations Seasons



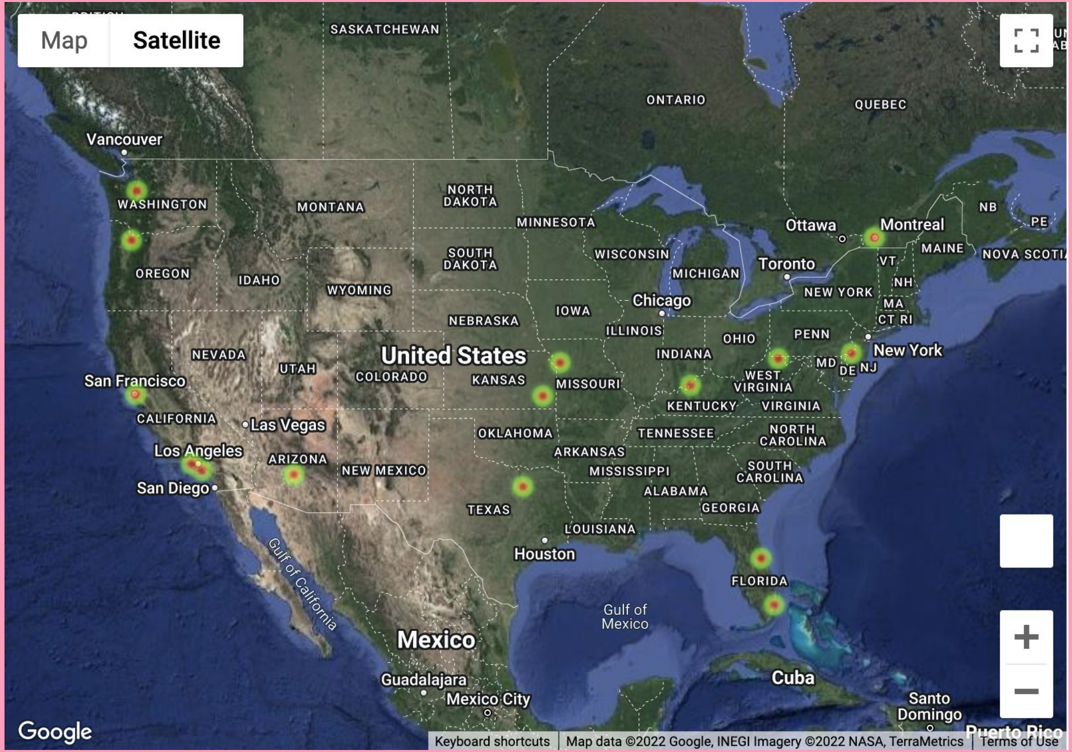
Contestants That Made The Top 5





	Finalists	Contestants	Percent
Massachusetts	3	9	0.333333
Colorado	3	11	0.272727
Oregon	2	8	0.250000
Washington	3	13	0.230769
North Carolina	2	10	0.200000
Texas	6	31	0.193548
Ohio	2	14	0.142857
Florida	4	32	0.125000
Tennessee	1	8	0.125000
New York	2	18	0.111111
Utah	1	11	0.090909
Arizona	1	11	0.090909
Michigan	1	12	0.083333
California	5	66	0.075758

Locations of Winners



Do Contestants from specific regions have a better chance of success?

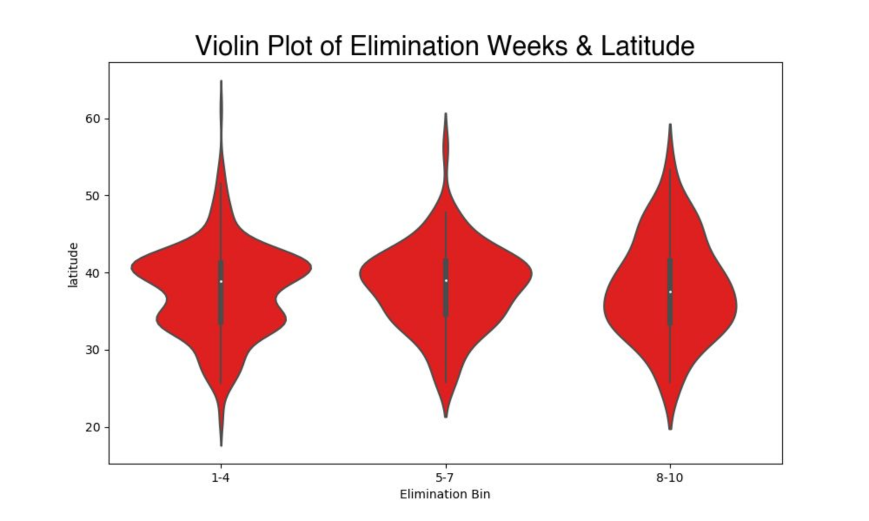
-Majority of the winners were from the southern region of the United States.

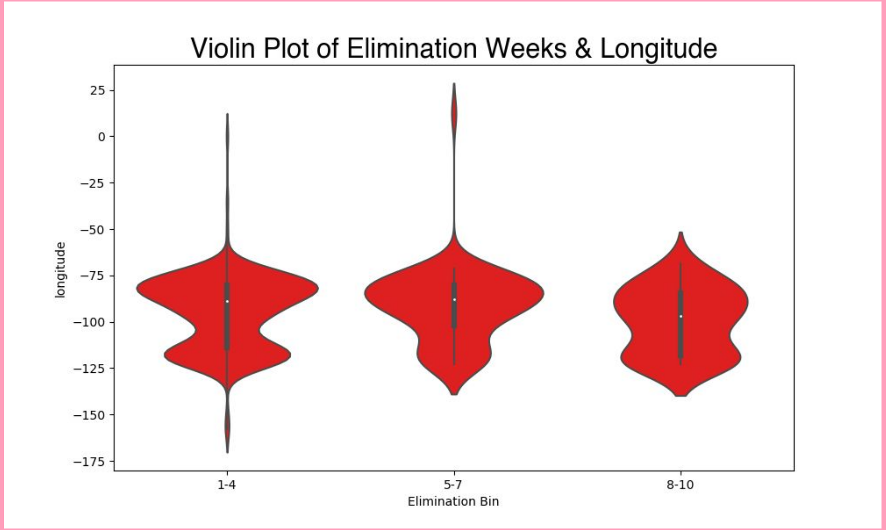
-Majority of contestants from all Seasons are from California.

-California had the lowest percentage of finalists at 7.5% and Massachusetts had the highest percentages of of finalists at 33%.

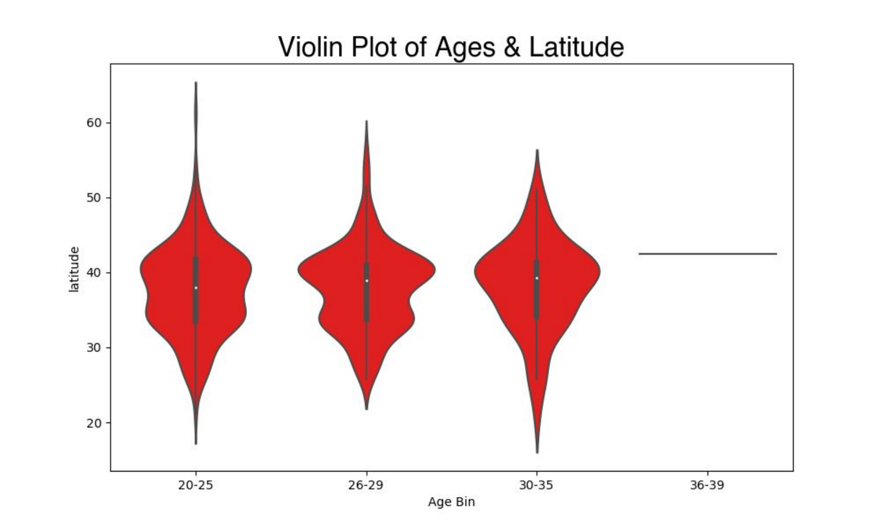
-Generally, contestants from the South have a better chance of success, however it is not a reliable variable to make a predictive model.

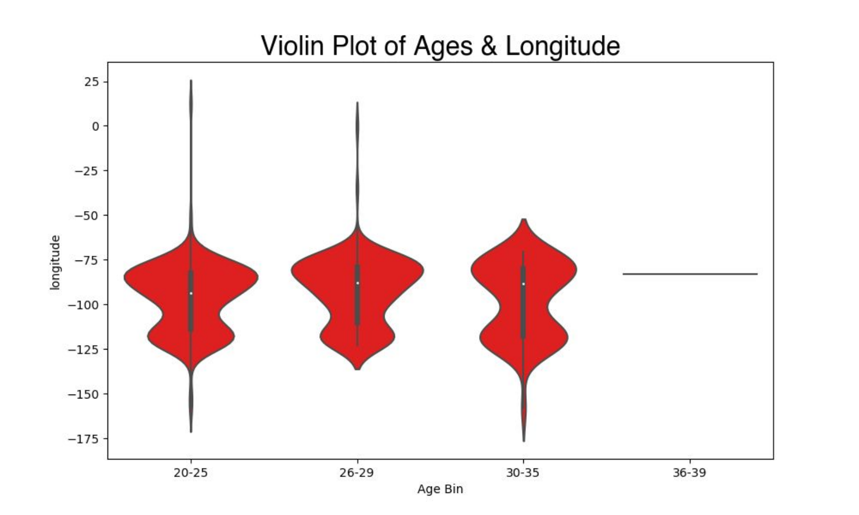
Violin Plots: Elimination Weeks & Latitude & Longitude



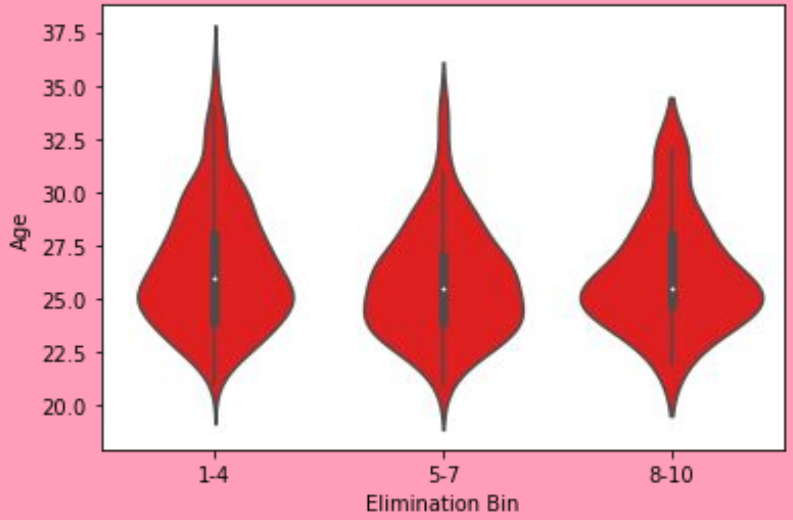


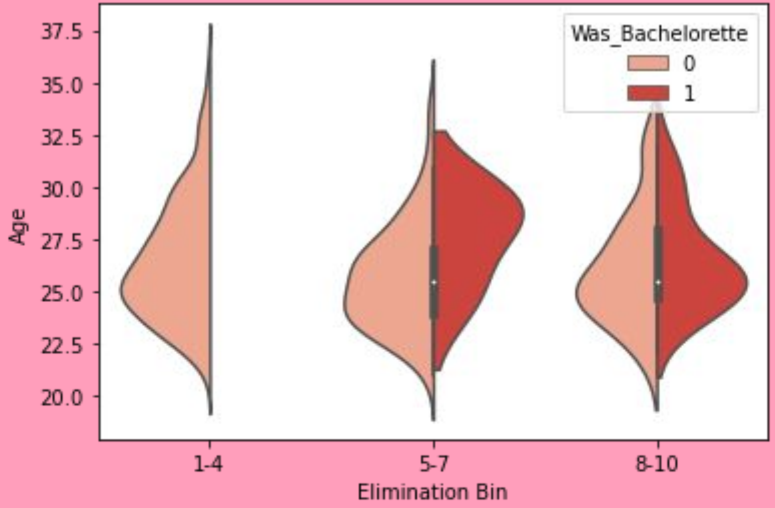
Violin Plots: Age Groups & Latitude & Longitude





Violin Plot: Elimination Week & Age





(As well as Bachelorette data)

Statistics & Regression

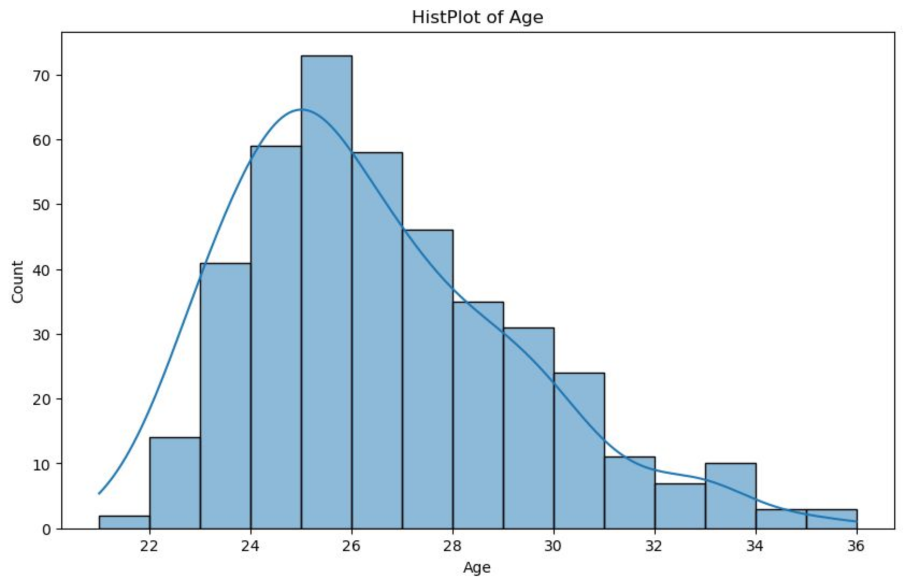
Does having the age, occupation, and home location of a contestant, allow us to make significant predictions on how well they will do while on *The Bachelor*?

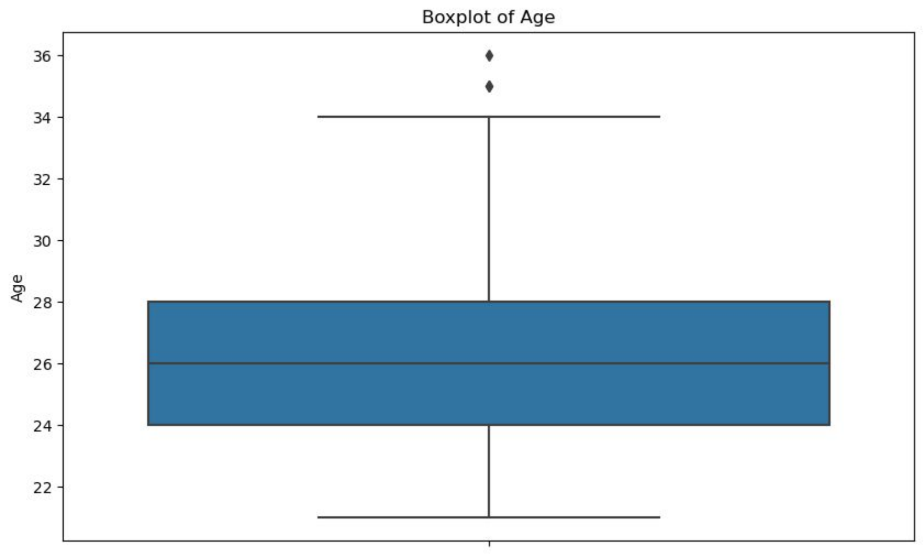
- Used “One Hot Encoding” technique to do a statistical summary, since we had categorical data about Occupations.
  - This technique changes the categorical values into numerical values, where each is represented as binary vectors, 1’s and 0’s.

- Our linear regression checked the shape of our data using histograms, and boxplots.

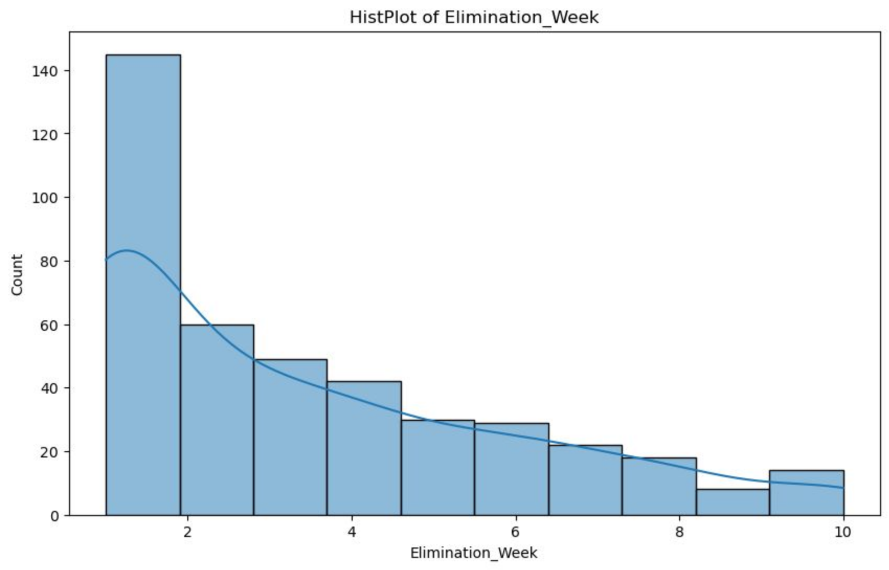
- We saw that our data was not normally distributed, had outliers, and that our model would most likely not be correlated.

Histogram and Boxplot of Age Data

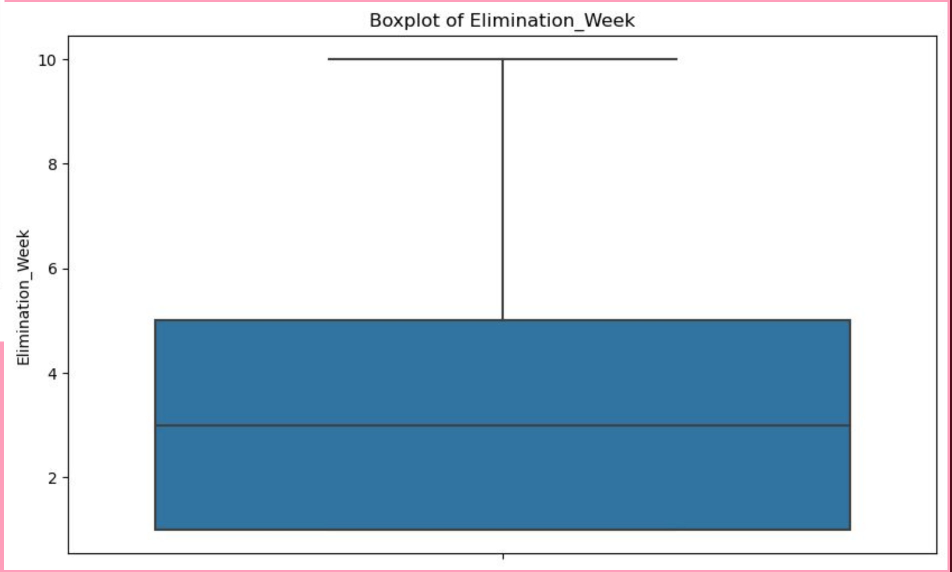




Histogram and Boxplot of Elimination Data

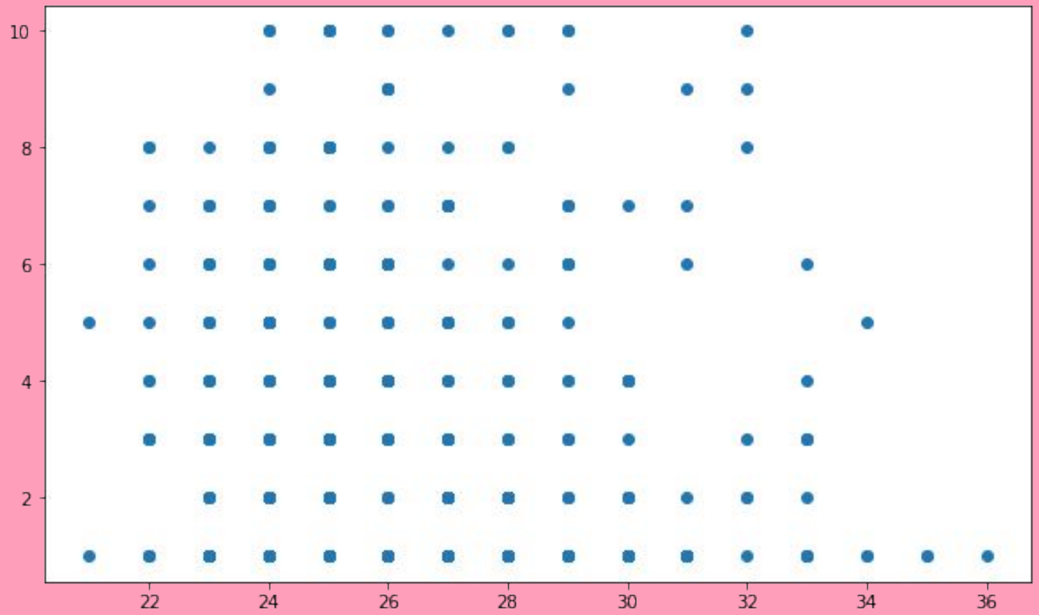


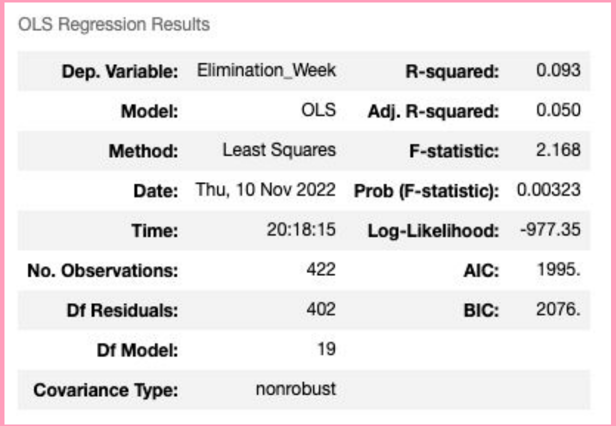
HistPlot of Elimination_Week



Boxplot of Elimination_Week

Scatter Plot of Age & Elimination Week



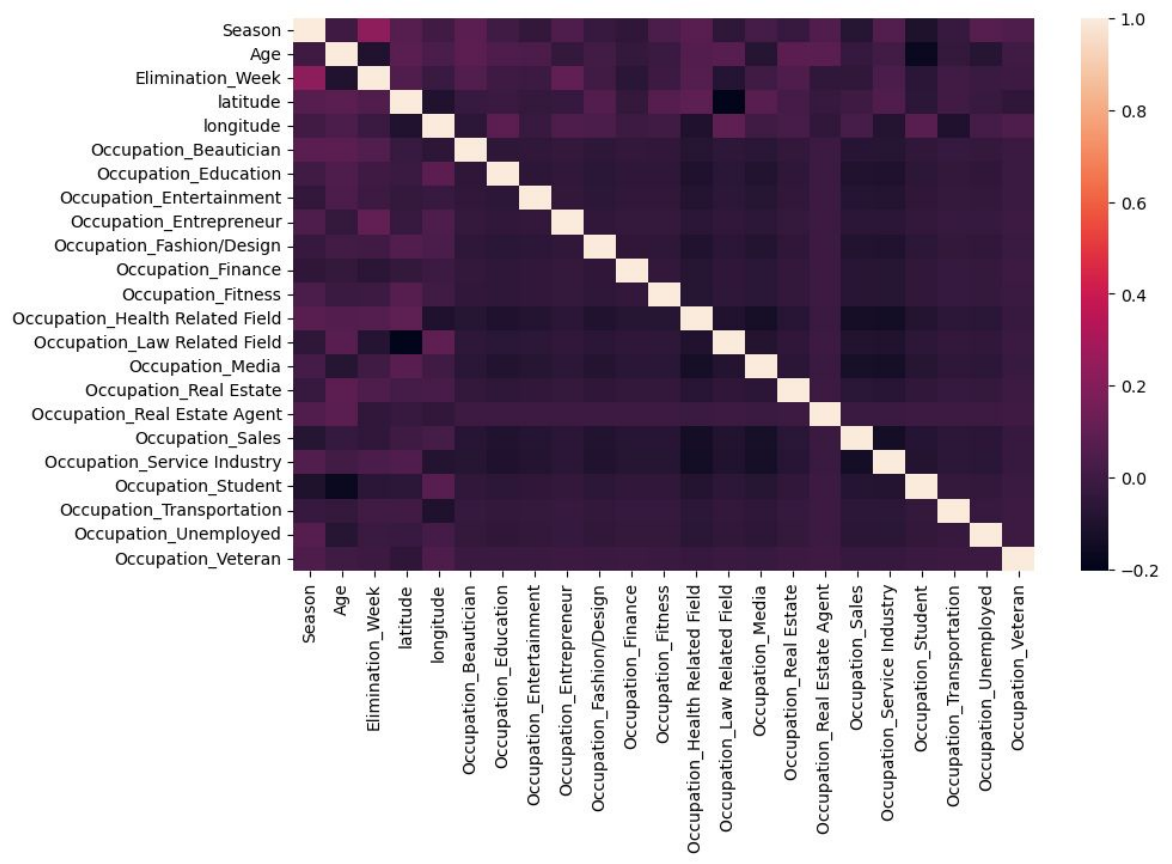


OLS Regression Results

Dep. Variable:	Elimination_Week	**R-squared:**	0.093
Model:	OLS	**Adj. R-squared:**	0.050
Method:	Least Squares	**F-statistic:**	2.168
Date:	Thu, 10 Nov 2022	**Prob (F-statistic):**	0.00323
Time:	20:18:15	**Log-Likelihood:**	-977.35
No. Observations:	422	**AIC:**	1995.
Df Residuals:	402	**BIC:**	2076.
Df Model:	19		
Covariance Type:	nonrobust		

Both the Scatter Plot & R-Squared value indicates no correlation

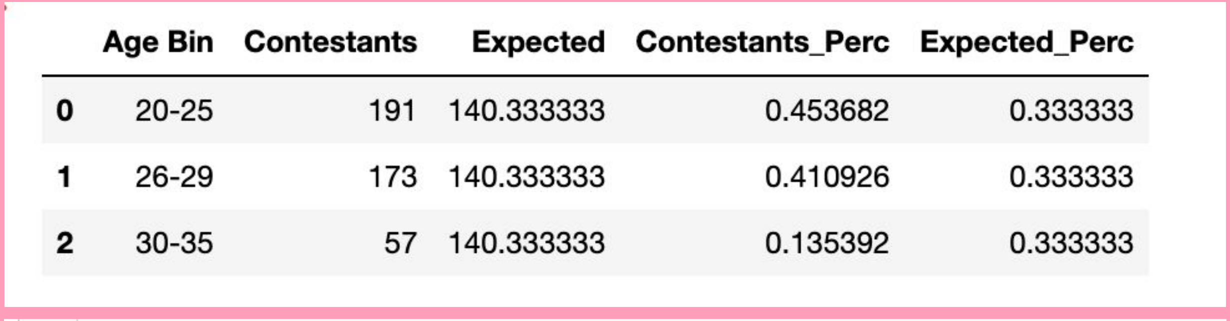
Correlation Heatmap



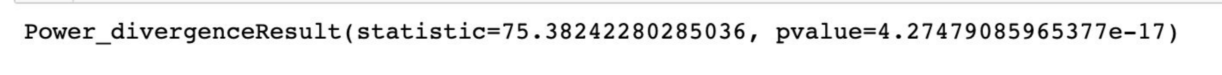
Chi Square Test

Null Hypothesis:
There is no difference between the distribution of age groups of contestants and the bachelor is fair.

Alternative Hypothesis:
There is a difference between the distribution of age groups of contestants and the bachelor is not fair.



	Age Bin	Contestants	Expected	Contestants_Perc	Expected_Perc
0	20-25	191	140.333333	0.453682	0.333333
1	26-29	173	140.333333	0.410926	0.333333
2	30-35	57	140.333333	0.135392	0.333333



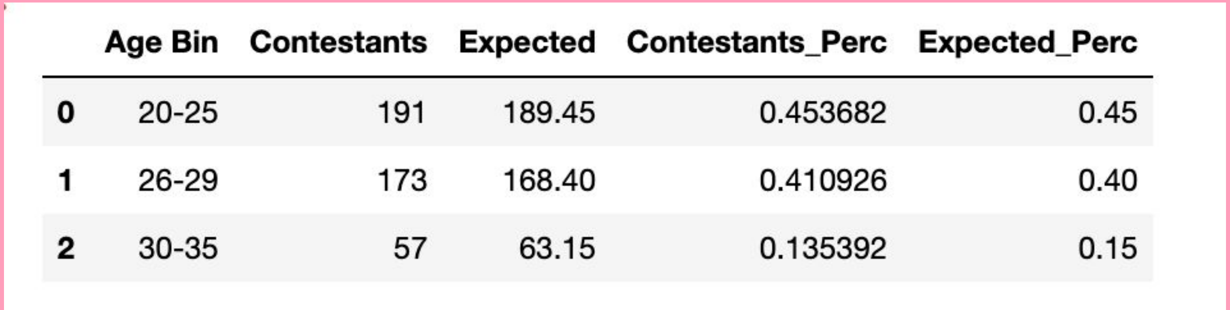
```
Power_divergenceResult(statistic=75.38242280285036, pvalue=4.27479085965377e-17)
```

We reject the null hypothesis because of the p-value

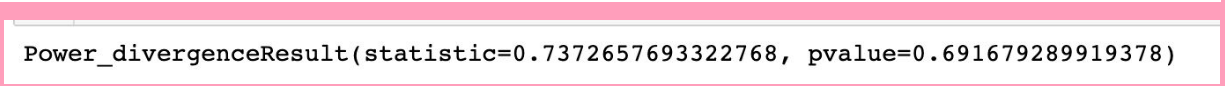
Chi Square Test

Null Hypothesis:
There is no difference between the distribution of age groups of contestants and the bachelor is fair.

Alternative Hypothesis:
There is a difference between the distribution of age groups of contestants and the bachelor is not fair.



	Age Bin	Contestants	Expected	Contestants_Perc	Expected_Perc
0	20-25	191	189.45	0.453682	0.45
1	26-29	173	168.40	0.410926	0.40
2	30-35	57	63.15	0.135392	0.15



```
Power_divergenceResult(statistic=0.7372657693322768, pvalue=0.691679289919378)
```

We fail to reject the null hypothesis because of the p-value

Limitations

Dataset Limitations:

- Not all Seasons were represented from the given datasets. Contestant information contained information only from Seasons 1, 2, 5, 9-21. Missing Seasons 3, 4, 6-8.
- No information regarding race to make demographics more specific.
- Occupation information was extremely diverse.
- Since it is a reality television show, some predictions are nearly impossible to predict given that it is produced to create entertainment, drama, ratings, etc., and there are a number of other factors that aren't easily measured (emotions, feelings, behind the scenes information) that could lead to a person's chance of success.

Conclusion

Conclusion

- Our prediction model wasn’t strong and we can not make a complete correlation between our variables.
- Age range has gotten younger over the shows lifetime.
- The amount of contestants increased over time.
- Location analysis shows that contestants from the Southern region of the United States have a better chance of success on the show, but location is not a reliable variable to make a prediction.
- Occupations have remained fairly consistent throughout all the seasons, and through the elimination process showing that there is no correlation with a contestants occupation and how far they make it on the Bachelor.

Alexander Booth Chances of Finding Love on The Bachelor

Name: Alexander Booth

Occupation: Senior Data Analyst

State:Texas

City: Dallas

Age:30’s

Chances of Success: Even with living in Dallas Texas, as that does well in the show. There is not one single Data Analyst occupation in the data and with the shows trend for going younger,it is likely the chances of finding love on the show is low

Works Cited

https://en.wikipedia.org/wiki/The_Bachelor_(American_TV_series)

https://www.kaggle.com/datasets/brianbgonz/the-bachelorette-contestants

https://www.kaggle.com/datasets/rachelleperez/the-bachelor-vs-the-bachelorette?select=contestants.csv

https://data.world/amandanovak/bachelor-contestants-with-instagram-follower-count

https://openweathermap.org/

https://developers.google.com/maps

https://coolors.co/palette/641220-6e1423-85182a-a11d33-a71e34-b21e35-bd1f36-c71f37-da1e37-e01e37

https://coolors.co/palette/ff0a54-ff477e-ff5c8a-ff7096-ff85a1-ff99ac-fbb1bd-f9bec7-f7cad0-fae0e4

https://coolors.co/palette/590d22-800f2f-a4133c-c9184a-ff4d6d-ff758f-ff8fa3-ffb3c1-ffccd5-fff0f3

https://coolors.co/palette/ffe0e9-ffc2d4-ff9ebb-ff7aa2-e05780-b9375e-8a2846-602437-522e38

https://bouqs.com/blog/50-official-state-flowers/

https://www.google.com/imghp?hl=en&ogbl

https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/#:~:text=A%20one%20hot%20encoding%20is,is%20marked%20with%20a%201.