The Bachelor Exploratory Data Analysis Write Up

Project 1 - Group 5

Matthew Bailey
Alexandra Flores
Will Hughes
Vanessa Martinez

Introduction

Our first exploratory data analysis project was on the popular reality TV show *The Bachelor*. *The Bachelor* is an American dating and relationship reality television series that debuted on March 25, 2002, on ABC. The show revolves around a single bachelor who begins with a group 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 episode, the bachelor interacts with the women contestants and presents a rose to each woman he wishes to remain on the show. Those who don't receive a rose are eliminated.

We chose to do our first exploratory data analysis project on *The Bachelor* to try and gain insight about contestant demographics, and to see if we would be able to successfully make predictions regarding a contestant's success on the show. We focused on previous contestants' age, the amount of time they were on the show, their occupation, and where they were from, to conduct our analysis.

Data Information and Cleaning

To start our analysis, we used two separate datasets on *The Bachelor* from Kaggle.com that contained csv files and an excel spreadsheet from Data.world. In Jupyter Notebook, we imported Pandas, numpy, matplotlib.pyplot, in order to do the bulk of our analysis. We had three files we needed to merge together. Two from Kaggle and one from Data.world. This was Will's time working with zip files so he got to learn how to upload those to his notebook. Once everything was loaded in, the bulk of the work on the data cleaning side was getting rid of all the repeat columns. We ended up dropping 13 of the original 19 columns. Fortunately most of the data was already in a very workable place when we started so we didn't have a lot of null values to deal with. The datasets included information on contestants that was narrowed down to the following: Contestant Name, Age, Occupation, Hometown, Seasons of Show, Elimination Week of Contestant, and Place (Rank). Seasons, Ages, and Elimination weeks were grouped together using bins.

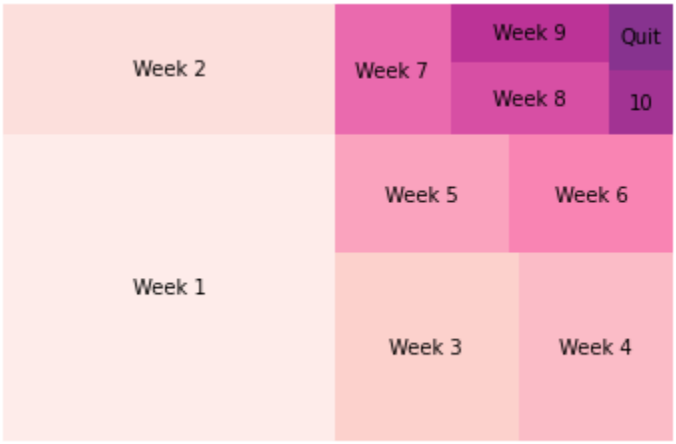
For location information, importing json, requests, time, and Openweathermap.org API geocoding was used to assist with finding latitude and longitude of the contestants' given home location of City, State and Country. To create cleaner information about location, rows that had null cities were dropped and rows that had null State values were filled in with a space. A contestant address was created by combining, City, State, and Country. For all addresses in the data set, a loop was made to request latitude, longitude, City, State, and Country to openweathermap.org, and then merged to the original cleaned data frame. Google Cloud/Maps was then able to be used to generate maps showing various locations of contestants. For Statistics information and visualizations, scipy stats and seaborn were imported into the notebook. For linear modeling we used statsmodels.api, linregress from scipy.stats, LinerRegression from sklearn.linear_model, and RandomForestRegressor from sklearn.ensemble.

Research Questions & Analysis

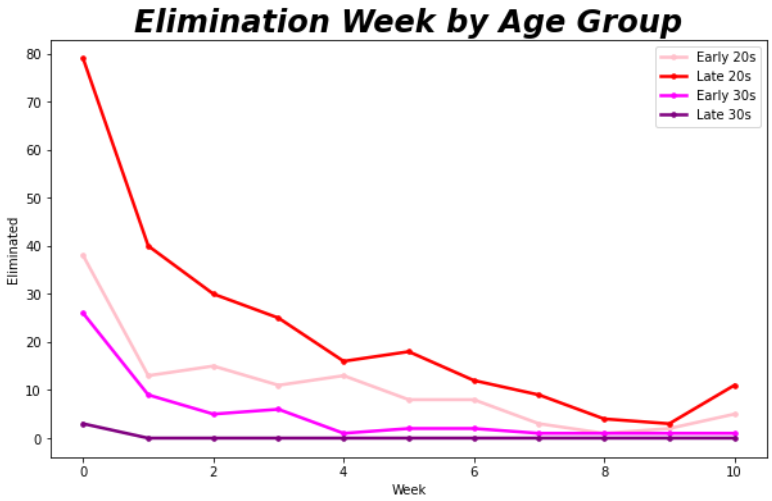
To make an analysis based on this data, we asked four research questions and divided them up between each team member. What was the average elimination week per age range? Has the age range and individual age of contestants changed over time? Do specific occupations lead to a contestant making it longer into the courtship? Do contestants from specific regions have a better chance of success?

Elimination Week Analysis

I took a look at the eliminations by week and, as predicted, found that the vast majority of eliminations happen at the beginning of the season.

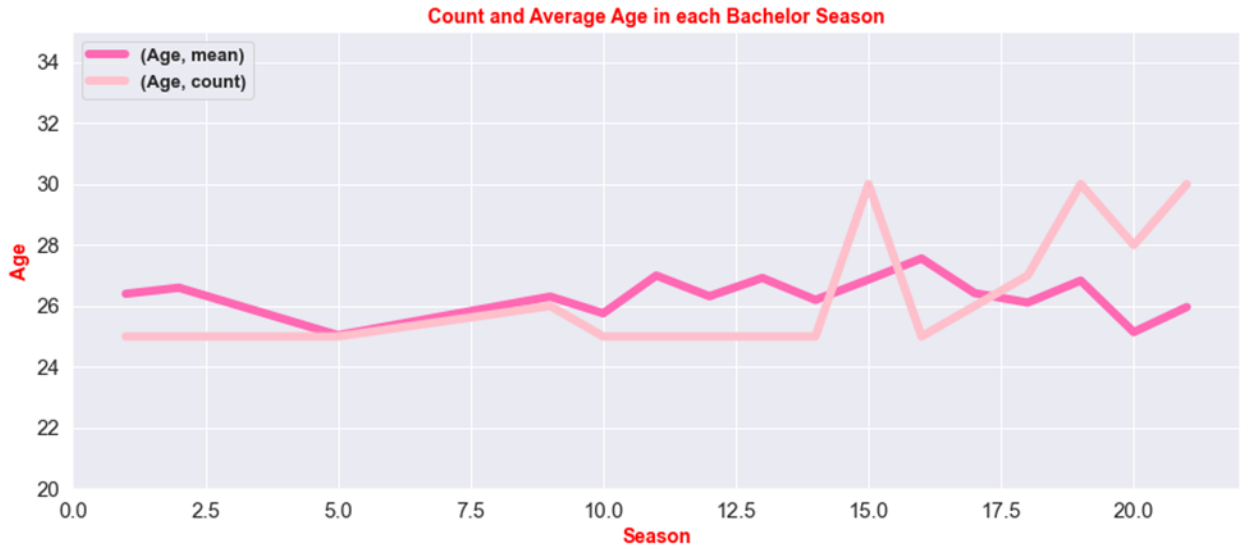


Just under half of all eliminations happen in the first two weeks, and decrease as the show progresses (and the pool of contestants is whittled down). I also showed that because of the higher probability of elimination in the beginning of the season, and the higher percentage of contestants in their early 20's, that more young contestants are eliminated than old.

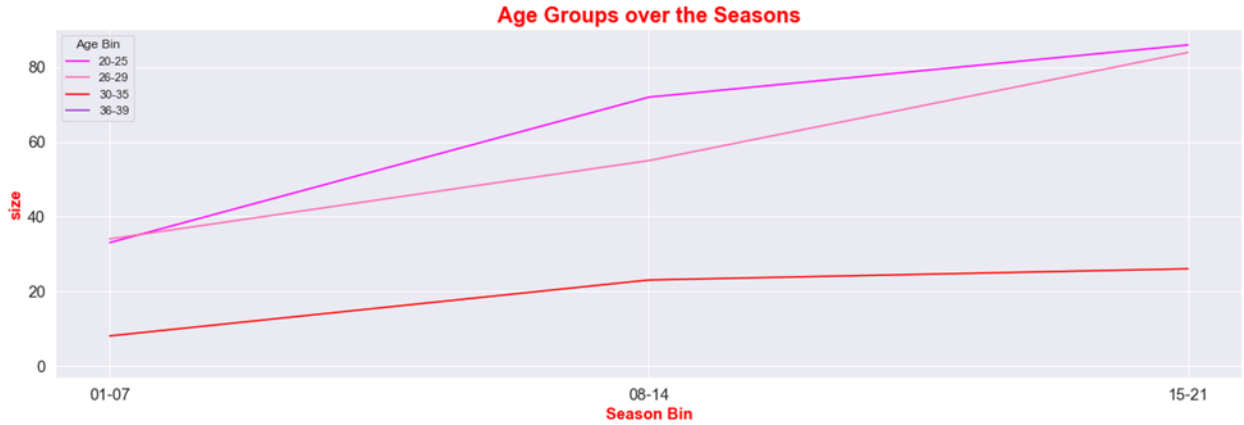


Age Analysis

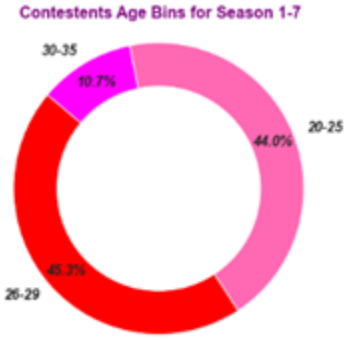
My first step was to find the average age of the contestant in the show and to see if there was a trend of the average age moving up or down throughout the show's lifetime. My next step was to find the best data visualization for this. After going through several visualizations, the easiest to read was a line graph.

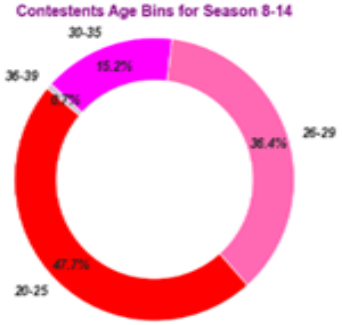


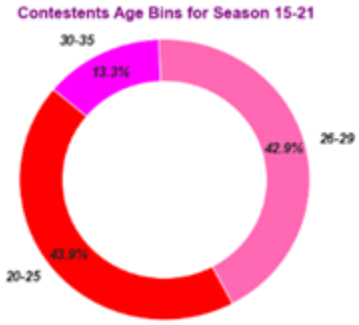
Once I completed the line graph, I then moved on to answer the question of whether the age range changed over time and if there were any correlations. The first step of this was to organize the contestants in "Age Bins" and the seasons in "Season Bins". Organizing the seasons and contestants in bins would give me the ability to see if the show trended younger or older as time went on.

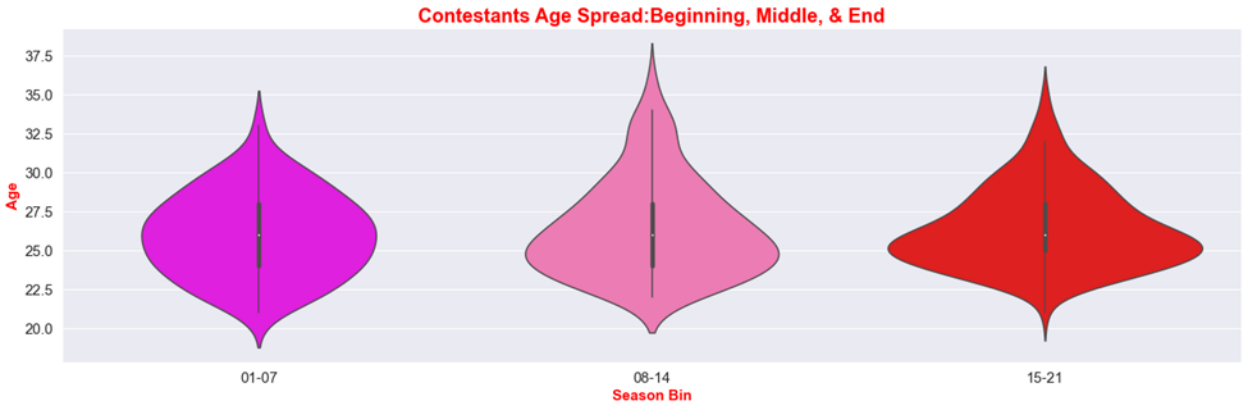


To illustrate this, three kinds of visualizations became apparent as the best to use. These were a line graph, a violin chart, and a donut chart.





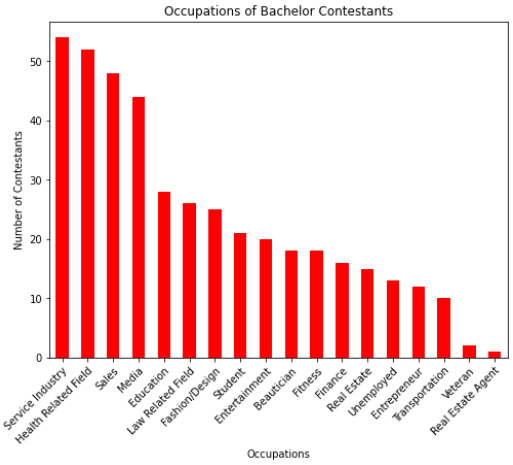




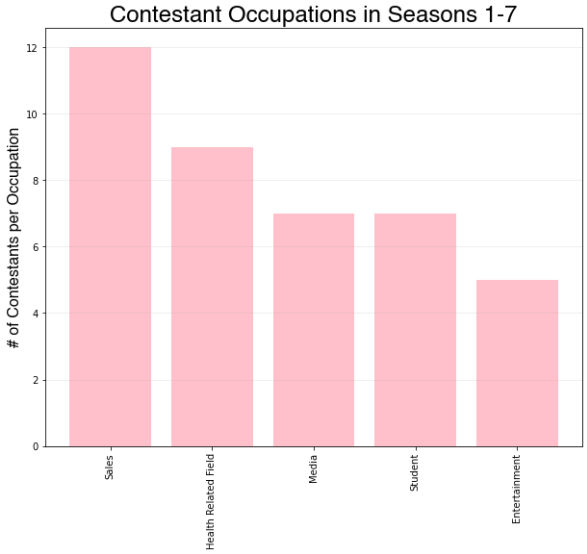
These visualizations helped us see that throughout the 21 years of The Bachelor the ages have continued to vary and the proportion of these age ranges have not significantly changed. If the data set was complete and included a few more columns such as viewership numbers or the age of the Bachelor in each season, there would be many questions we could possibly answer. Such as what was the average age of the contestant in comparison to the age of the Bachelor? Did that age trend older or younger in the later seasons of the show? Did the age ranges of contestants get younger or older as the show went into the later seasons? Did this correlate with viewership numbers?

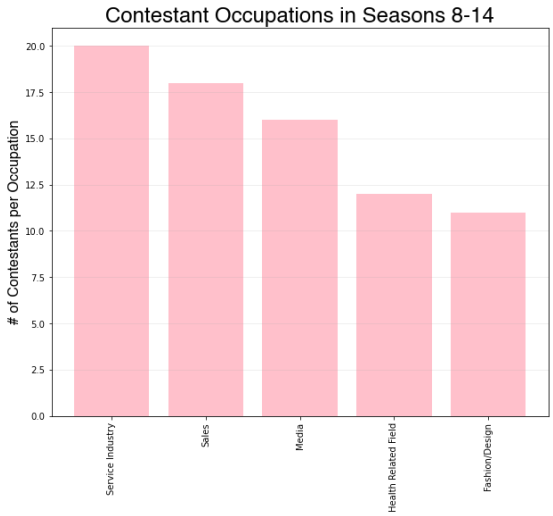
Occupation Analysis

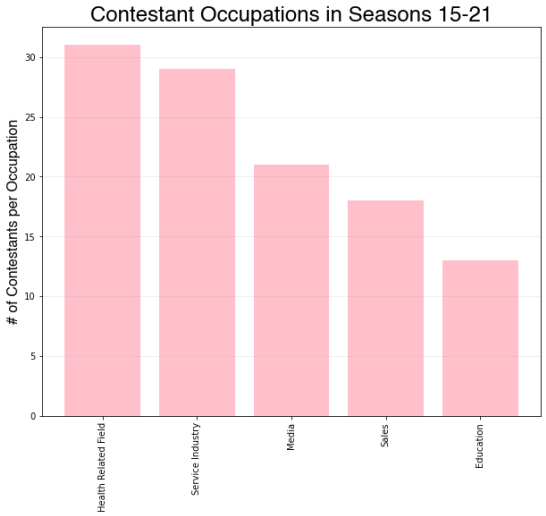
The raw occupation data for the bachelor contestants had a great variety as there were 424 contestants. In order to make an analysis we had to create parent categories for all of them resulting in 18 different categories.



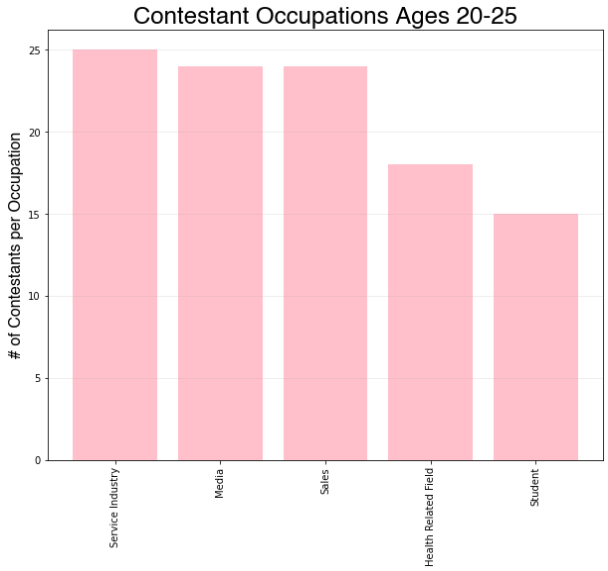
To look into the data even further we also grouped the data by seasons 1-7, 8-14, 15-21, elimination week, and we also grouped the occupations by age ranges.

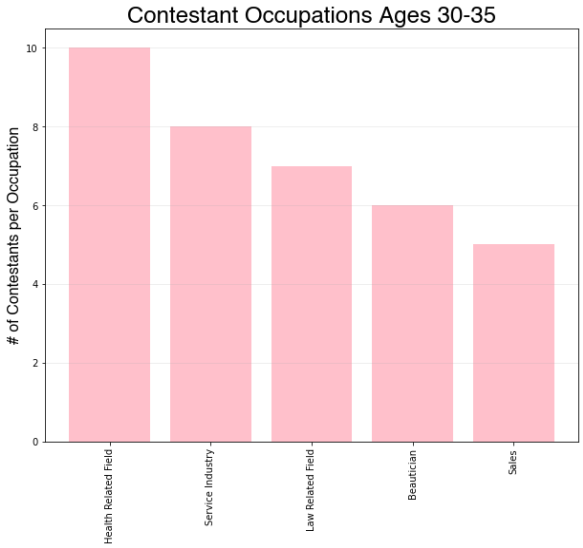


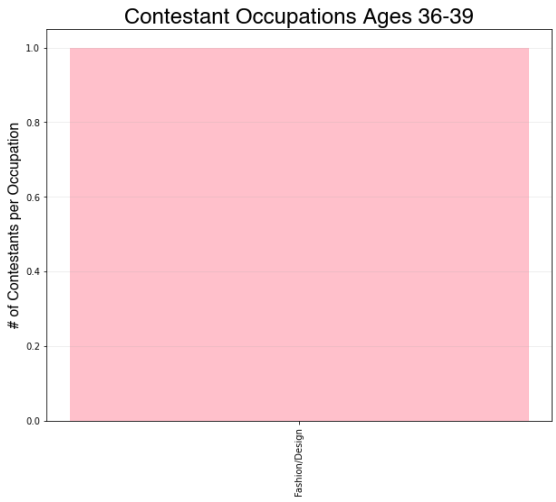




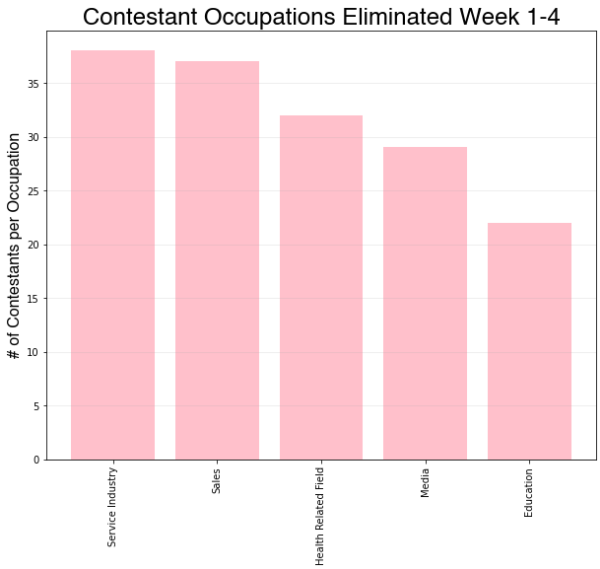
After going over the occupations by season ranges the data remained pretty consistent with the most common occupations being health related, service industry, media, and sales. This set of data told us that contestant occupations have not changed much since 2003. So we then looked into our age range data that categorized the occupation of the contestant by their age 20-25, 26-29, 30-35, 36-39.

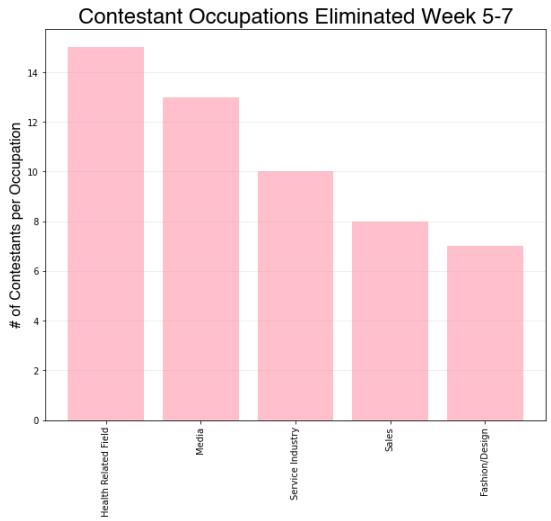


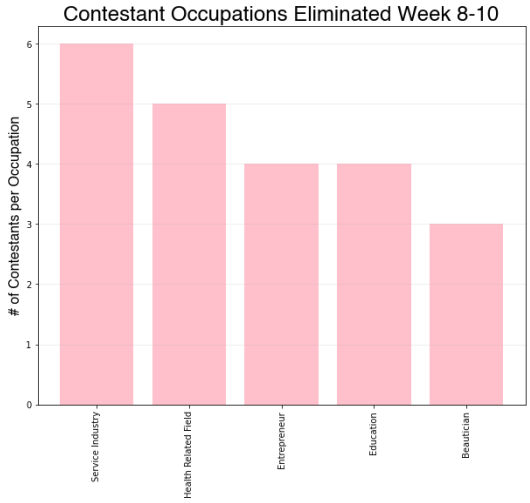




The data showed the same consistency as before with the most common occupations being service industry, health related, sales, and media, telling us that age did not make a difference regarding a contestant's occupation. So we then went through and grouped by elimination week to see if we could find any significant difference in occupation of contestants that make it farther along in the courtship.



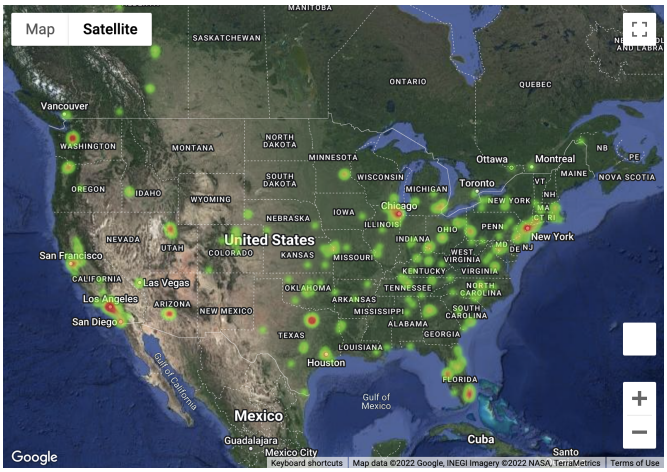


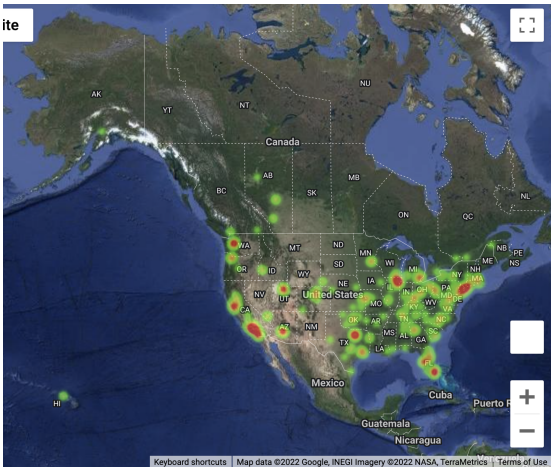


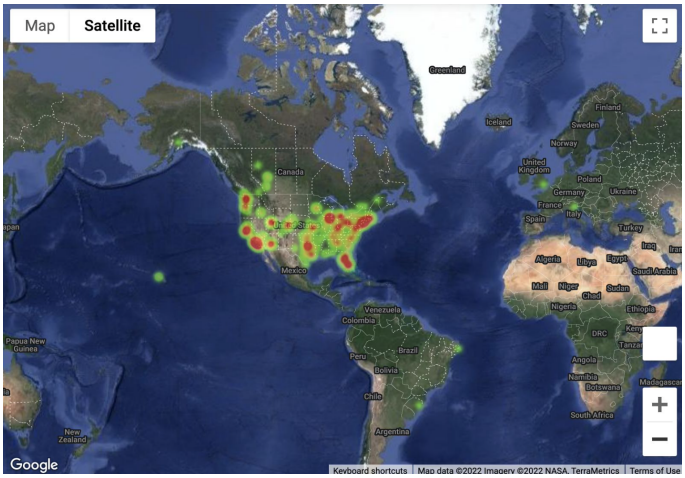
The data showed that health related occupations, service industry, and sales remained popular throughout the season elimination rounds. Since the same occupations were the most common throughout each elimination week, we concluded that occupation does not affect how far you make it along the courtship. Something we could've done to give more insight is grouping the occupations by location to see the relationship there and to compare the most common occupations in big cities versus small ones.

Location Analysis

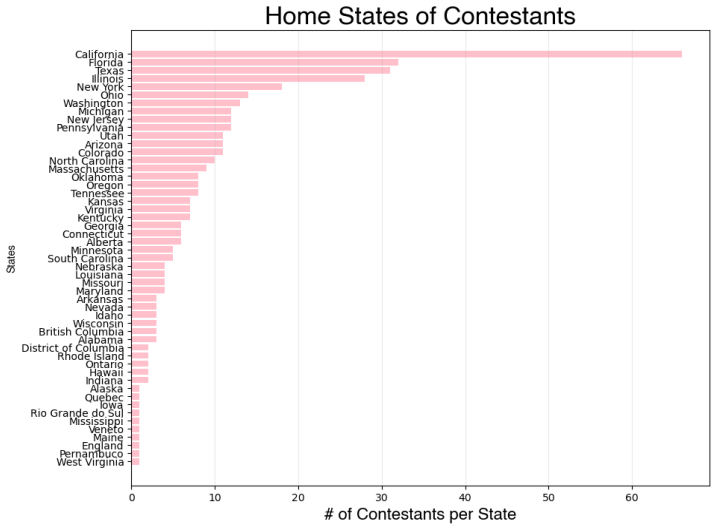
Contestants' home location could be made into clear visualizations by using the latitude and longitude coordinates. Heatmaps from Google maps showed the locations of all contestants from all the seasons in our dataset. Majority were from the United States and North America, and some were from South America and Europe.



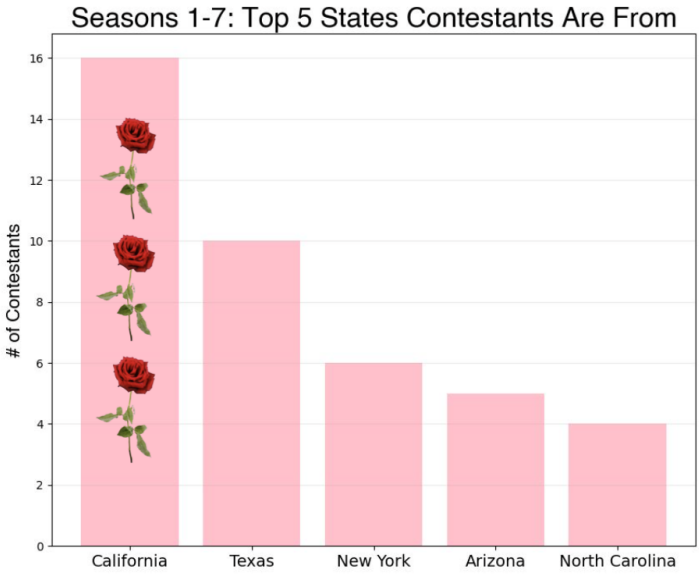


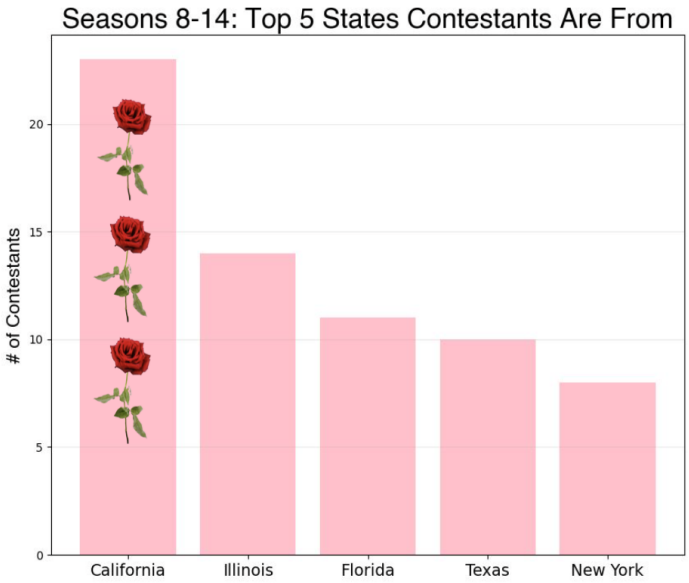


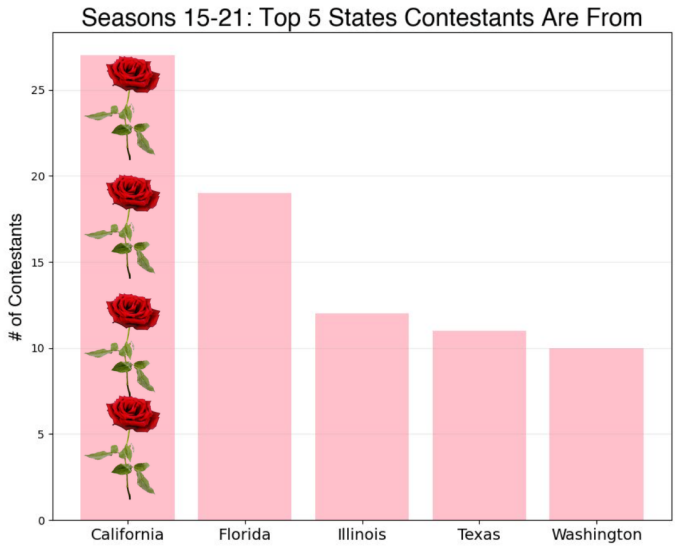
Bar charts were made to show the distribution of contestants and where they were from, across various seasons of the show. This showed that the majority of contestants from all seasons were from California.



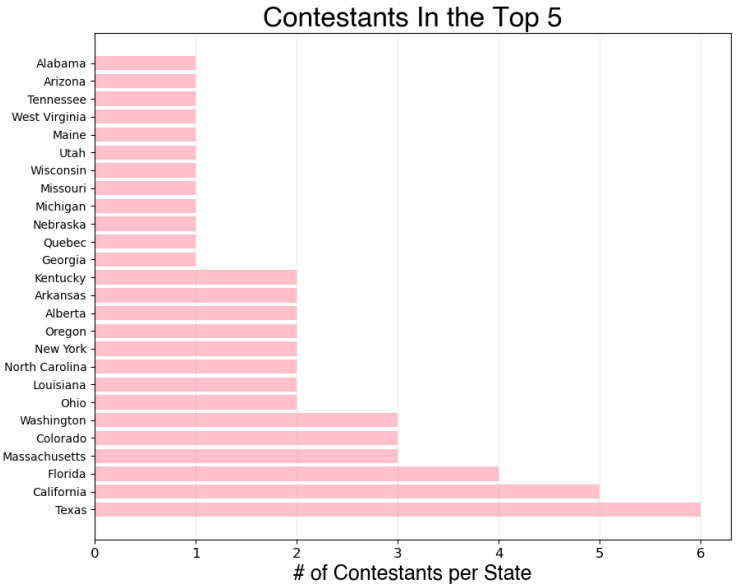
The top five States of where contestants were from included California, Florida, Texas, Illinois, and New York. For seasons one through seven, California, Texas, New York, Arizona, and North Carolina were the top five States. Seasons eight through fourteen showed California, Illinois, Florida, Texas, and New York were the top five States. Seasons fifteen though twenty-one showed California, Florida, Illinois, Texas and Washington were the top five States. Majority of contestants, throughout all seasons, were from the Southern region of the United States. However the Northeast, Midwest, and Northwest region were slightly represented throughout the course of the show.





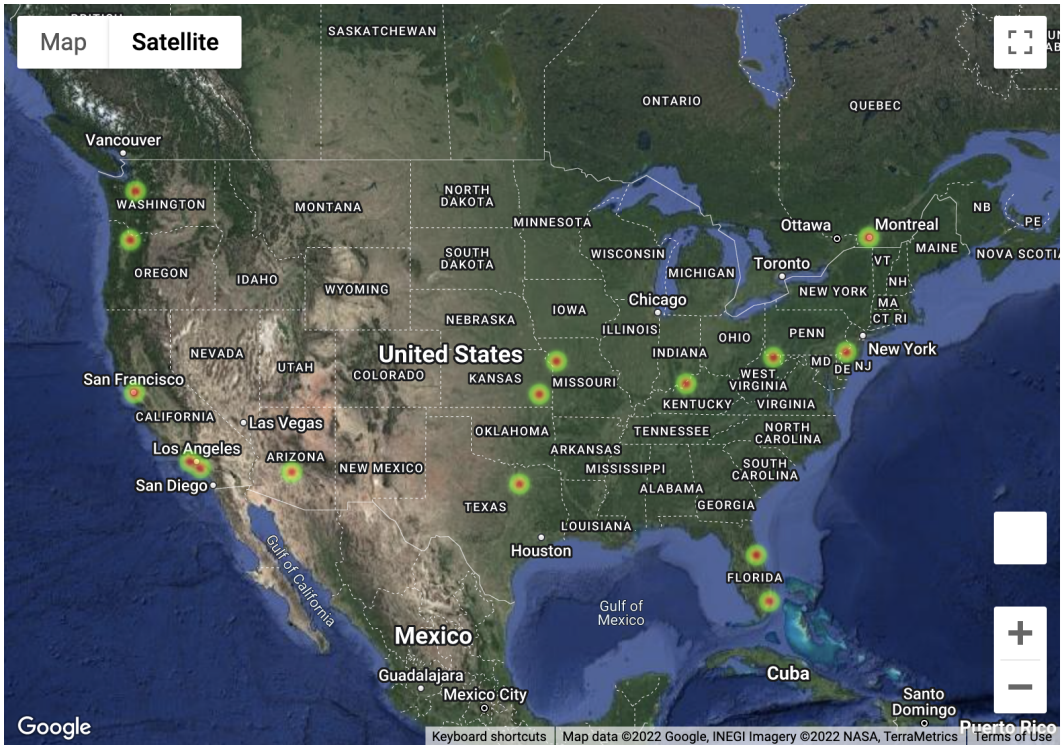


The States that had the greatest number of contestants make it to the top five in the show were Texas, California, Florida, Massachusetts, Colorado and Washington. Massachusetts produced the highest percentages of finalists at 33%, and California had the lowest percentage of 7.5%.



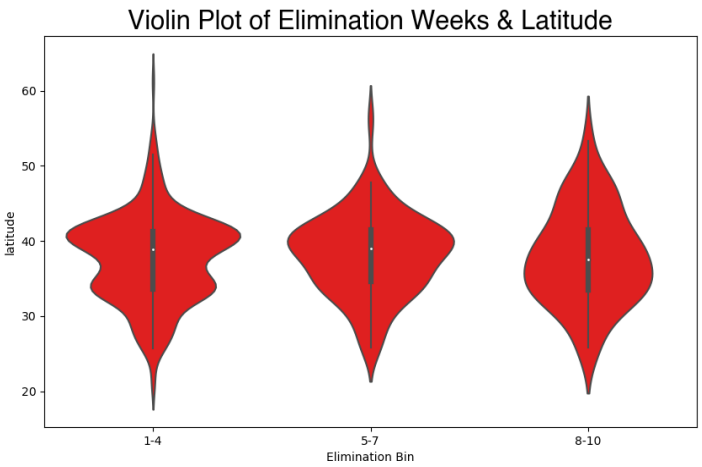
	Finalists	Contestants	Percent
Massachusetts	3	9	0.333333
Colorado	3	11	0.272727
Oregon	2	8	0.250000
Washington	3	13	0.230769
North Carolina	2	10	0.200000
Texas	6	31	0.193548
Ohio	2	14	0.142857
Florida	4	32	0.125000
Tennessee	1	8	0.125000
New York	2	18	0.111111
Utah	1	11	0.090909
Arizona	1	11	0.090909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

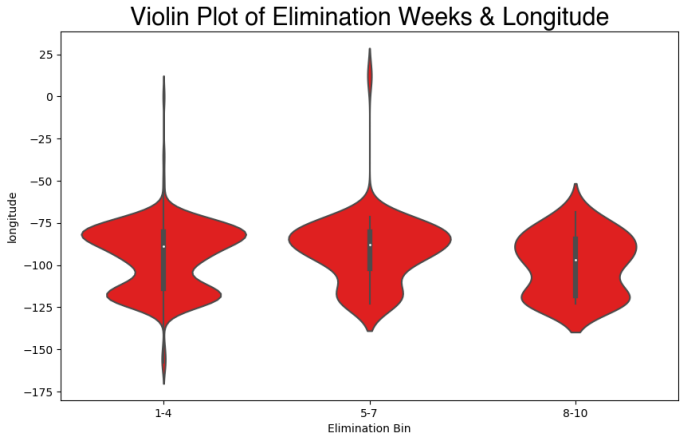
Contestants that won *The Bachelor* were from Arizona, California, Florida, Kansas, Kentucky, Missouri, New Jersey, Oregon, Quebec, Texas, Washington, and West Virginia.



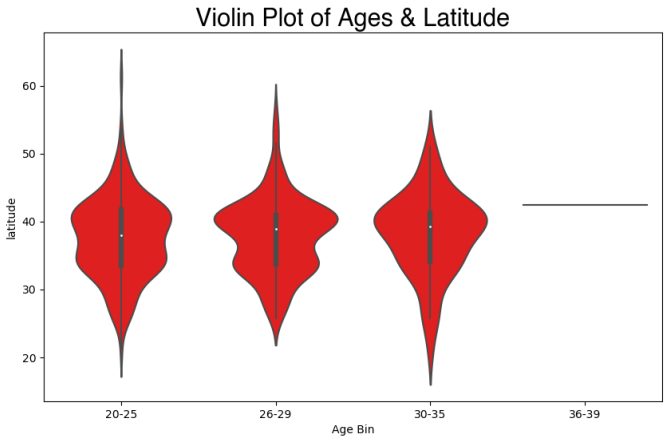
Season	clean_state
1	Kansas
2	New Jersey
5	California
9	Florida
10	California
12	California
13	Texas
14	Florida
15	West Virginia
16	Arizona
17	Washington
18	Missouri
19	Kentucky
20	Oregon
21	Quebec

Violin Plots of grouped elimination weeks and latitude/longitude showed that contestants from Southern latitudes were more represented, and eliminated in later weeks, compared to contestants from the North. There was not a pattern between elimination weeks and longitude because the contestants from the Southern region do well, and balance out the contestants from Northern regions who do poorly.





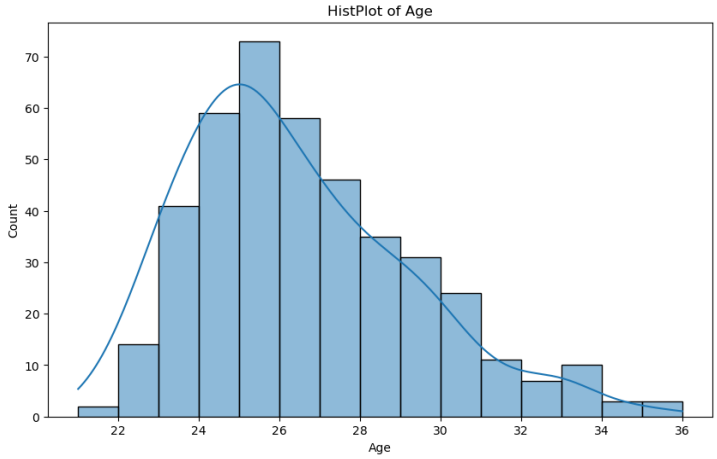
Violin Plots of grouped ages and latitude had a pattern showing that younger contestants, in the 20-25 age group, were more likely to be from the South compared to any other age group.

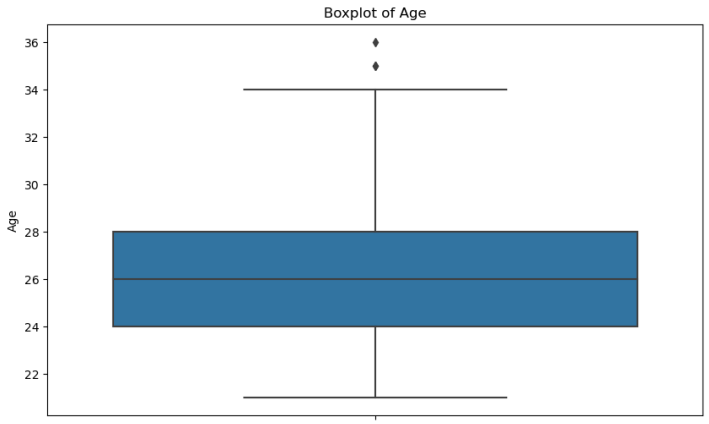


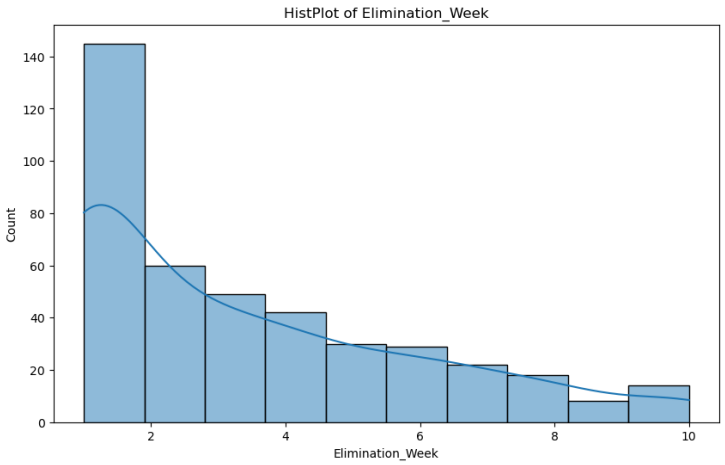
In summary, the majority of the winners were from the Southern region of the United States and the majority of contestants, from all seasons, were from California. California, however, had the lowest percentage of finalists at 7.5% and Massachusetts had the highest percentages of finalists at 33%. Generally, based on location, contestants from the South have a better chance of success on *The Bachelor*, but it is not a reliable variable to make a predictive model.

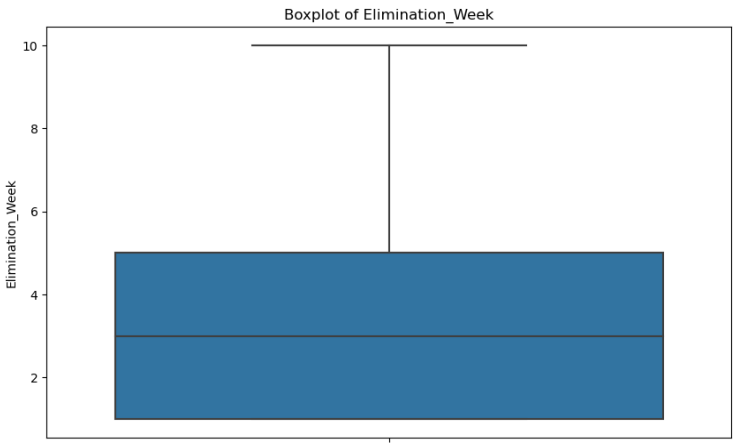
Regression

For our regression we asked, does having the age, occupation, and home location of a contestant allow us to make significant predictions on how well they will do, or how long they will last, while on *The Bachelor*? Our linear regression checked the shape of our data using histograms, box plots, and a scatter plot. We saw that our data was not normally distributed, had outliers, and that our model would most likely not be correlated. We used the "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.

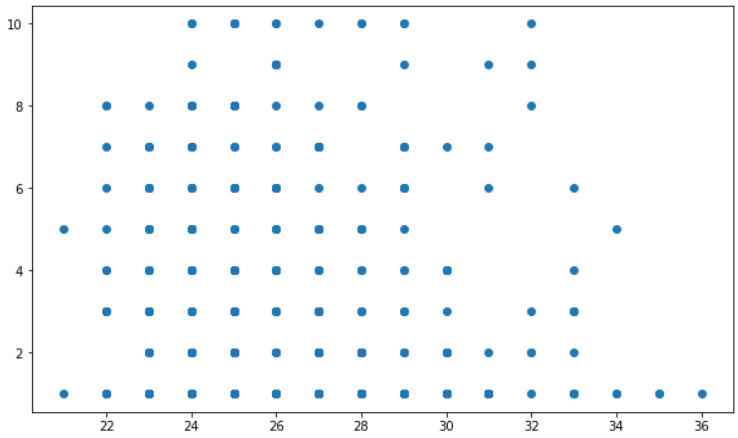








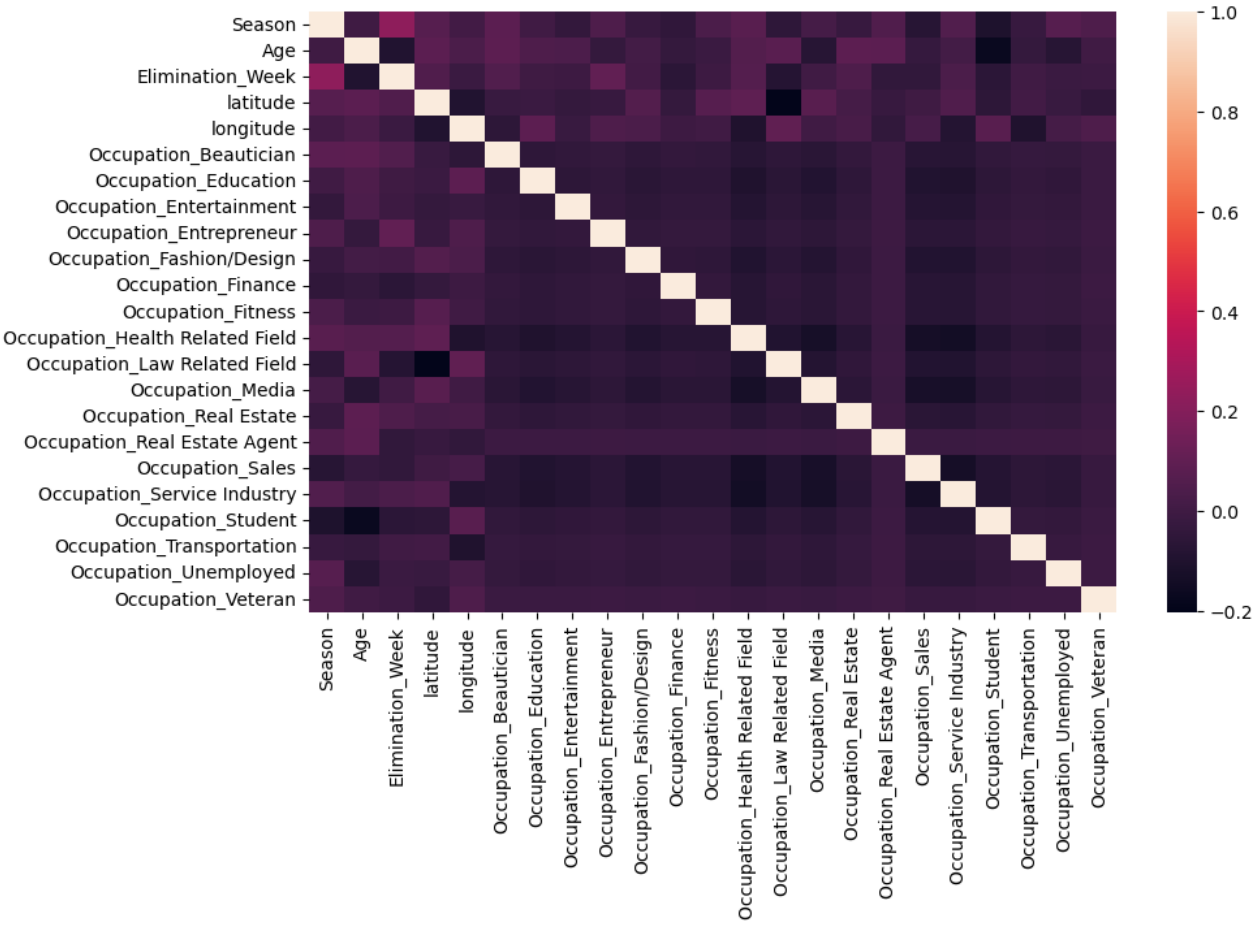
The Scatter Plot shows Age and Elimination Week. Both the Scatter Plot and R-squared value in the OLS Regression results, indicate there is no correlation.



OLS Regression Results

Dep. Variable:	Elimination_Week	**R-squared:**	0.092
Model:	OLS	**Adj. R-squared:**	0.044
Method:	Least Squares	**F-statistic:**	1.902
Date:	Sun, 13 Nov 2022	**Prob (F-statistic):**	0.00995
Time:	17:34:11	**Log-Likelihood:**	-966.81
No. Observations:	417	**AIC:**	1978.
Df Residuals:	395	**BIC:**	2066.
Df Model:	21		
Covariance Type:	nonrobust		

The correlation heatmap also shows the same.



Chi-Square Test

A Chi-Square Test was done regarding the observed data for contestant ages. Our null hypothesis was that there is no difference between age bins/groups, and *The Bachelor* is fair when determining contestants by age. Our alternative hypothesis was that there is a difference between age bins/groups, and *The Bachelor* is not fair when determining contestants by age. The first test was to see if there was an equal distribution between the three age groups, including ages 20-25, 26-29, and 30-35. The test showed an extremely low p-value, so the null hypothesis was rejected.

	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

```
# Run the chi square test with stats.chisquare()
st.chisquare(df7.Contestants, df7.Expected)
```

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

The second test was to see if the distribution of the three age groups of contestants was weighted differently, specifically the youngest age group of 20-25 at 45%, ages 26-29 at 40%, and ages 30-35 at 15%. The test showed a p-value of 0.69, so the null hypothesis was not rejected. Typically, a p-value greater than 0.05 means that deviation from the null hypothesis is not statistically significant, and the null hypothesis is not rejected.

	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

```
st.chisquare(df7.Contestants, df7.Expected)
```

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

Conclusions

Conclusions of our analysis show that our prediction model wasn't strong and we could not make a complete correlation between our variables. What was apparent was that the average age of the individual contestant did not change too much even though the number of contestants increased and the percentage of contestants in their 20s increased in the later seasons compared to the earlier seasons. Occupation analysis showed that occupations have remained fairly consistent throughout all of the seasons, and through the elimination process showing that there is no correlation with a contestant's occupation and how far they make it on *The Bachelor*. 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. Our Chi-Square Test showed that there was not an equal distribution between the age groups of contestants chosen to be on the show. Thus, the producers of *The Bachelor* most likely want the age distribution to be closer to 45% in the 20-25 age range, 40% in the 26-29 age range, and 10% in the 30-35 age range .

Limitations & Bias

There were limitations in our dataset that affected our analysis. We did not have all of the seasons of *The Bachelor* provided from the given datasets. The contestant information contained only seasons 1, 2, 5, and 9-21. We were missing seasons 3, 4, 6-8. There was no information regarding race, to make the contestant demographics more specific. The occupation information that was provided 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 including emotions, feelings, and behind the scenes information that could lead to a person's chance of success. Because of these reasons, the ability to find correlations and trends in age was more difficult as well.

Future Work

Future work for this analysis could be done using more information from the seasons that we did not have in our dataset. We could see if there is data regarding other demographics of contestants, including race and religion. An analysis could be done to show how long a relationship has lasted between the winning contestant and the bachelor, after the show has ended for the season. An analysis can also be made regarding fan support and how long a contestant lasts in the courtship, as we know the fan base and producers can impact how long a contestant stays in the running. There is a large following of fans for this show, and there is even an account on Instagram called "bachelordata" that shows various data analytic information for *The Bachelor* and its spin off shows. Until the show stops airing on television, there can be various kinds of data analysis done in order to tell a story about *The Bachelor* contestants.

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://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/
https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/#:~:text=A%20one%20hot%20encoding%20is,is%20marked%20with%20a%201
https://www.google.com/imghp?hl=en&ogbl
https://mashable.com/article/bachelordata-bachelor-nation-interview