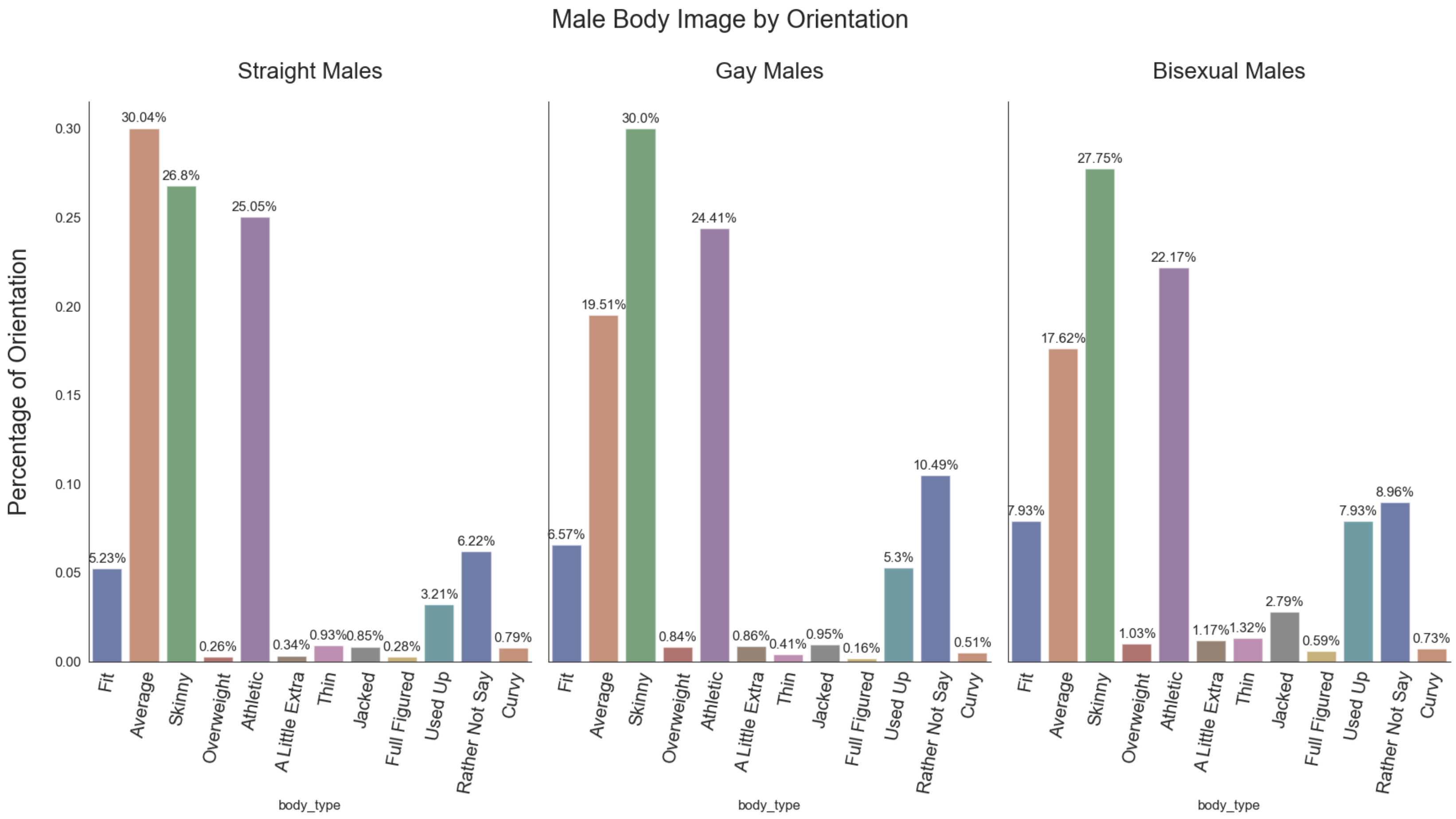
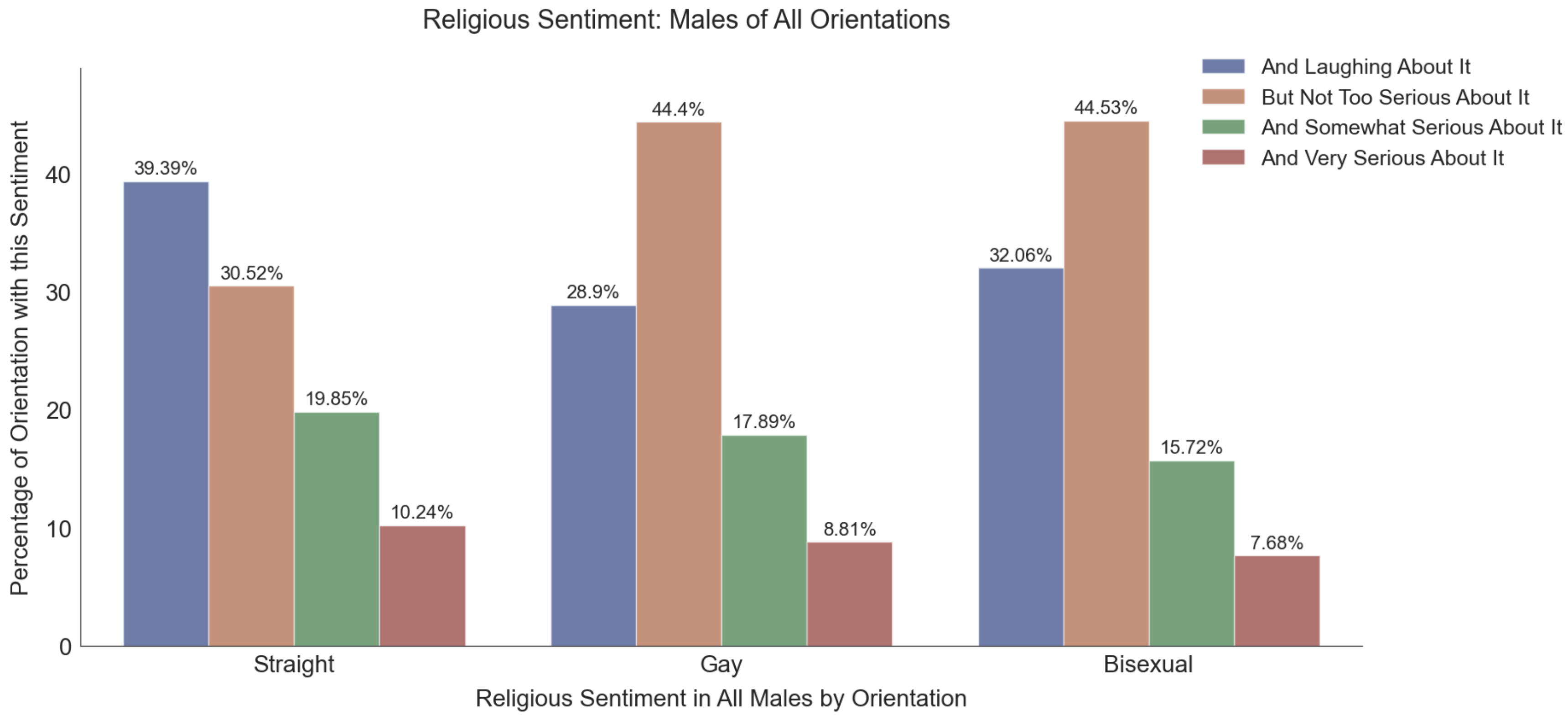
Predicting male orientation

IS IT POSSIBLE? TO WHAT DEGREE?

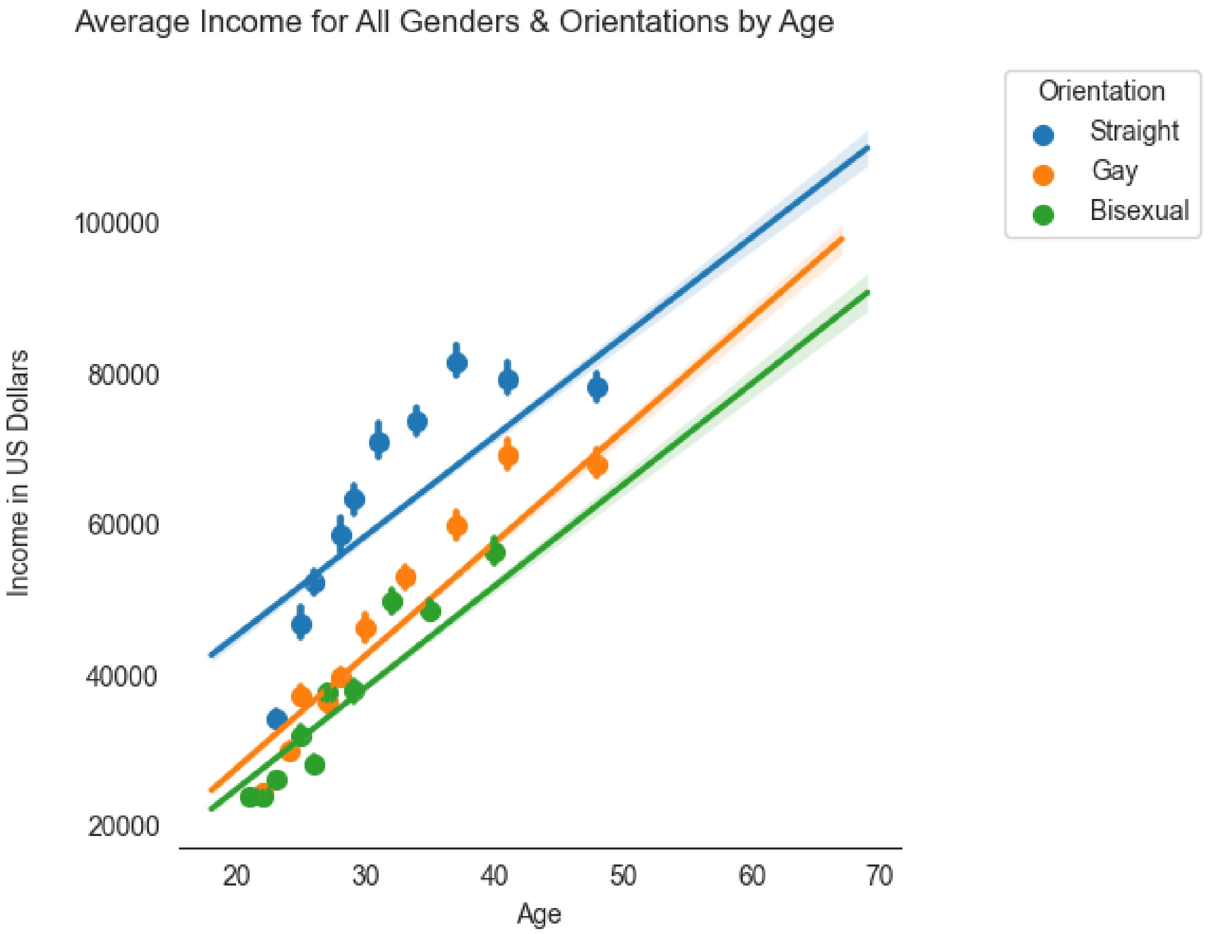
Graph One: Nested Bar Charts with Labeled Columns Male Body Type per Orientation



Graph Two: Grouped Bar Charts with Labeled Columns Male Orientation & Religious Sentiment



Graph Three: Seaborn Multiple Regression Plot Average Income by Age & Orientation



Primary Questions & Purpose

- **What Questions Did I Seek to Answer?**

1. What is the relationship between ‘primary’ ethnicity and orientation?
2. Is there a relationship between orientation and average income?
3. What is the average salary in each professional sector for each orientation?
4. Does generational status (e.g., Baby Boomers vs. Gen Z) impact orientation in males?
5. Does orientation influence substance abuse?

- **Why Did I Choose to Focus on Orientation?**
- A dating site is the one place people would not lie about sexual orientation.
- People will fudge on other numbers. They may pad age, income and height. But it defeats the purpose of a dating site to lie about who you want to date. It seemed like a rare opportunity to explore how orientation relates to other factors, including traditional markers of social and professional success.

Explanation of New Columns

- I created new columns for new data in several ways.
- I created a “city,” column from the location with pd.apply and lambda functions using ‘rsplit’ to separate city from state.
- I used ‘explode’, ‘expand’ and ‘split’ to separate religious affiliation from religious sentiment to create a ‘sentiment’ column.
- I used ‘pd.cut’ to separate males by calculated birth years into generations: Silent Generation, Baby Boomers, Generation X, Millennials and Generation Z.
- I used ‘split’ and ‘insert’ to parse zodiac sign from sentiment about the zodiac, and to prepare the features_data dataframe for mapping.
- I also used the conventional device of equating the new data to a new column in the designated dataframe, as needed.

0.09511
0.09511

Comparison Between Two Classification Approaches

- The non-linear K Nearest Neighbors and Decision Tree classifiers produced the best results. I ran them using only balanced classes and other factors, including ethnicity,
- The results did not inspire great confidence that one can determine male orientation based on statistical differences in the three classes from the dataset.
- The Decision Tree classifier was simpler. It returned the highest r-squared value (89%) and a much faster speed (Runtime: .0911 seconds). It also gave me the opportunity to experiment with predictive classifications for orientation based on fictional profiles.
- That opportunity showed sociological variables (e.g., city and religion) played the decisive roles in predicting male orientation, far above genetic factors like ethnicity.
- The Decision Tree's accuracy was 69%. Precision and recall – quality vs. quantity - were also comparable, both 68% as weighted averages, which validates the accuracy score.
- The K Neighbors classifier was much slower (34 seconds). Though it's r-squared score was only 38%, it's accuracy score was higher than the Decision Tree at 85%, with precision and recall scores at 82% and 80% for weighted averages.

Comparison Between Two Regression Approaches Ethnicity and Male Orientation

- I searched for linear equations to describe the relationship between ethnicity and male orientation, comparing the results from Multinomial Linear Regression and Binary Logistic Regression.
- Both were simple to implement. The binary, logistic regression ran twice as fast (.067) compared to the linear regression model (`0.16495` secs), running both models with both imbalanced and balanced classes.
- The independent coefficients for the linear regression changed depending on the classes used, but the r-squared value remained the same at ~5%, which means class imbalance did not explain some differences.
- The low, sometimes negative, coefficients reject that ethnicity and male orientation share a linear relationship.
- The results improved in the logistic regression, with an r-squared value of 48% - much higher than the linear regression. The accuracy score (also 48%) indicates the model does not have much predictive power.
- Again, the weighted average for precision (47%) is the same as the accuracy score. The weighted average for recall (43%) is five points lower. It appears to confirm the r-squared score, meaning we found a low rate of success using the logistic model to assess male orientation and ethnicity (alone).
- These values do not change when we adjust the parameters.
- In addition to the discussed models, I also tried the Naïve Bayes Classifier, Support Vector Machines with different kernels, and the Random Forest, all of which validated the differences in orientation and ethnicity in the charts was due to widely varying sample sizes for each ethnicity.

Conclusion Statement

1. Non-linear classification models fared better than linear regression methods, but the statistical differences, for almost every topic, were not strong enough to make predictions about male orientation, whether I resampled and balanced the classes or not.
2. But I found some interesting takeaways:
3. Thirteen percent (13%) of males did not identify as straight, which challenges current estimates of 6.7% for the entire LGBT community in the San Francisco area, where 97% of males in the dataset live. Location (city) was a critical factor in orientation.
4. Females are 2.5 times more likely to identify as bisexual than males, while males are 2.5 more times more likely to identify as gay.
5. Twenty-three (23.4%) of straight men were somewhat to very serious about religion, compared to 26.7% of gay men, and 30.08% among bisexual men.
6. Religion was another important factor. Straight men were far more likely to identify with traditional religions, while gay/bisexual gravitated to non-traditional and Eastern religions.

Conclusion Statement

Major takeaways, continued.

1. The average straight income was more than twice the average income of gay males, though I suspect the issue lies partly with sample size.
2. 46-48% of gay and straight males had earned an undergraduate degree compared to 38% of bisexuals. However, gay/ bisexual males were more likely to earn Master's and doctorates.
3. Though gay and straight males had similar dietary preferences (79% and 82% for 'anything') bisexual males were far more likely to be vegetarians and vegans.
4. Forty-one percent (40.78%) of gay males reported not wanting children, compared to 31.67% for straight males and 10.11% for bisexual males.
5. The same percentage (7-8%) of both gay and straight males reported wanting to have children in the future, compared to 13.36% of bisexual males.
6. **There is an 11% drop in males who identify as straight from the Silent Generation to Generation Z, making age and generational status another defining variable.**
7. The number of gay males has increased by 5.5% during the same period, and the percentage of bisexuals has increased from .8 to 6.4% - or eight times higher.

Next Steps, etc…

- If I could start over, I would not focus on orientation, as there are few correlations greater than 2-3%. I would choose an approach that returned more compelling results.
- For my next step, if I continued with orientation, I would combine gay and bisexual male profiles, and compare two classes: straight and 'not straight."
- Most models seemed designed to work best with binary classes. This approach would have also reduced the problem of imbalanced classes, which was a major source of confusion.
- If I continued with this approach, I would hope for data compiled in confidential questionnaires versus a public dating site. The public nature and purpose of dating sites do not inspire honesty or depth.