About the Dataset

The dataset I'm working with contains detailed information about the top 195 billionaires from around the globe as of the year 2023. This comprehensive dataset encompasses various aspects of each billionaire's profile, including their ranking in terms of wealth, their current age, the final worth of their assets, as well as specific details such as their birth year, birth month, and birth day. Additionally, it categorizes each billionaire according to their specialized field or industry, providing valuable insights into the diverse sectors driving wealth accumulation at the highest levels.

Are we using PCA or not? Yes, because the first two PCs adds up to 90% which is greeater than 70%. Hence, we are using PCA for our analysis.

A graph with a line and a bar

Description automatically generated with medium confidence

Question 1: Can we classify individuals based on their demographic and financial attributes?

Answer:

Yes, we can classify individuals based on their demographic and financial attributes using logistic regression. By building a logistic regression model with variables such as age, final worth, and self-made status as predictors, we can predict an individual's gender with a certain level of accuracy. The model's coefficients indicate the direction and strength of the relationship between each predictor variable and gender.

Question 2: What are the main factors contributing to the variance in the dataset?

Answer:

Principal Component Analysis (PCA) helps identify the main factors contributing to the variance in the dataset. By examining the eigenvalues and eigenvectors, we find that the first few principal components explain a significant portion of the variance. These principal components represent linear combinations of the original variables, indicating which variables contribute most to the overall variance in the dataset.

A graph of numbers and words

Description automatically generated with medium confidence

As you can see the terms that are contributing to rank, age, finalWorth, birthyear, birthMonth and birthDay.

The code performs PCA on the dataset "b\_data" with different subsets of variables. The first PCA (b\_pca) considers only the age, final worth, birth year, and birth month, while the second PCA (b\_new\_pca) includes all variables except the rank. Scaling is applied in both cases, ensuring that each variable contributes equally to the analysis regardless of its scale.

Question 3: Can we predict an individual's final worth based on their demographic attributes?

Answer:

Using linear regression, we can predict an individual's final worth based on their demographic attributes such as age, birth year, and birth month. The regression coefficients indicate the direction and magnitude of the relationship between each demographic variable and final worth. By evaluating the model's performance metrics, such as R-squared and mean squared error, we can assess the accuracy of the predictions and the overall fit of the model.

A graph of a comparison between a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of

Description automatically generated

Here we are taking gender into consideration.

A well-performing model will have residuals scattered randomly around zero (the red dashed line). So, according to my understanding the female predicted vs actual is not matching whereas the male actual vs predicted is same.