# Predicting Minority Status of Boston Women-Owned Businesses

Arlyss Herzig, Anushka Pandey, Marco Hong, Deekshith Komira
DATA 200, Foundations of Data Analytics
Final Report
12/11/2023

## I. Introduction

Women have been historically placed at the margins of the economic sphere, excluded from opportunities for financial growth and equality. This form of gender-based financial ostracization is only exacerbated in minority communities, who are further excluded from formal financial networks.

Given this disparity, we plan to study the relationship between geographic location, type of business, and socioeconomic status of an area in relation to women-owned businesses, specifically women-owned businesses that are minority-owned, in the Greater Boston area. This research will help us understand these businesses' geographic clustering and if they are located in areas of relatively high or low socioeconomic status, with the goal of being able to provide policy makers with relevant information to promote the socioeconomic vitality of women, women of color, and their respective communities. In particular, if these businesses, and in particular minority-owned businesses, are in areas of low socioeconomic status, we can put economic resources toward supporting these businesses and communities where they may not have adequate access to financial resources.

Women-owned businesses tend to be predominantly service-oriented, retail, or consumer services [1]. Women face a disadvantage in running their own business, as they on average have substantially fewer years of work experience than men prior to starting [2], and often lack the financial skills necessary to have a successful business, largely due to structural and systemic inequities [2]. This disadvantage is exacerbated by the fact that women have difficulties with obtaining adequate money to finance their business [1].

Long-term business survival varies largely based on gender and race [3]. In general, male-owned businesses fare better than female-owned businesses, while Hispanic and Black-owned businesses fare worse than white-owned businesses [3]. When looking at intersectionalities of race and gender, it is clear that in most cases, businesses that are owned by women of color fare substantially worse than male or white-owned businesses [3].

The disparities prevalent between minority- and white-owned businesses are also due to the spatial barrier in which minority businesses that are located in majority-minority areas are limited in their connections to economic activity [4]. Because predominantly minority areas are often of lower socioeconomic status due to structural and systemic inequities, these businesses are often at an economic disadvantage [4].

To work toward gender and racial/ethnic economic equity, we must invest resources in women, immigrants, and people of color and their economic ventures. Supporting the economic growth of businesses of these historically marginalized groups will allow them to become financially

self-sufficient and raise their socioeconomic status while supporting the economy of the surrounding area through the introduction and continuation of businesses.

While past studies underscore the disadvantages faced by women- and minority-owned businesses, and therefore the need to provide them with adequate financial support, little attention has been paid to any type of predictive measure to determine if a business may need or be eligible for additional financial services or support. This gap likely remains due to the spatial variance of minority businesses and socioeconomic levels in different cities and regions. We aim to address this gap by looking at Boston specifically, with the goal of predicting if a business is women-owned or minority- and women-owned based on location, business type, and socioeconomic status of an area, and therefore be able to preemptively assess if a government can provide additional financial resources or support to ensure the business survival. We therefore present the following research question and hypotheses:

- Research question: How are business types, location, and socioeconomic status of the surrounding area correlated with whether a women-owned business is also minority owned?
- Hypothesis 1: Minority-owned businesses are clustered in areas of relatively lower socioeconomic status.
- Hypothesis 2: Service and retail businesses tend to be owned predominantly by minorities.

To address this research question, we begin with this introduction and literature review. The Methodology section explains the data source and pre-processing, summarizes exploratory data analysis (EDA) with data description and visualization of our two data sources, and then discusses our models we use to evaluate the data. The following section includes the results of our empirical analysis, and finally we conclude with a summary of the results of our study and the policy implications. Included after this are our references and the Appendix, which includes all figures from this paper and the knitted R Markdown file used to conduct these analyses.

## II. Methodology

Data: Source, Cleaning, and Pre-Processing

The data on women-owned businesses comes from the Boston city government [5] and contains information on the name and type of business, contact information, location, and if it is also owned by a minority, immigrant, or veteran – updated daily. To pre-process the data, we start by removing businesses (55 in total out of 251) in Excel that do not have a physical storefront or are no longer running, because we are looking at current businesses and the impact of their location. We then geocode the addresses in order to get coordinates in order to be able to join with our median income data, and then determine the distance and bearing from the center of Boston (defined as Government Center, 0° N).

To proxy for the socioeconomic status of an area, we spatially join the geocoded dataset with median income of Massachusetts census tracts from the American Community Survey 2021 5-Year Estimates (the most recent year available) [6]. We then remove any businesses (1 in total out of 195) that are located in areas with no residents and therefore no median income data, which are therefore not relevant to and cannot be used in our analysis. This leaves us with 195 non-duplicate businesses with associated median income of the surrounding census tract.

Finally, we recategorize the over 140 different unique business types into nine chosen and non-mutually exclusive categories – professional services, education, retail, creative economy, food, healthcare, beauty and wellness, services, and entertainment and culture – as dummy variables. We also create a dummy variable for if the business is minority owned or not.

Exploratory Data Analysis (EDA)

We begin by looking at the distribution of our variables – median income, business type, and median distance from Boston – by minority status to look for any underlying patterns in the data. Looking at business types, we found that education, retail, beauty and wellness, service, and entertainment businesses are owned substantially more by minorities (Figure 1). Next we see that when business types are further disaggregated by median income, minority businesses are overwhelmingly located in lower income areas across all business types (Figure 2). Lastly, when we disaggregate business types by median distance from Boston, we find that minority-owned businesses, except in education and beauty fields, are disproportionately located further from downtown Boston (Figure 3).

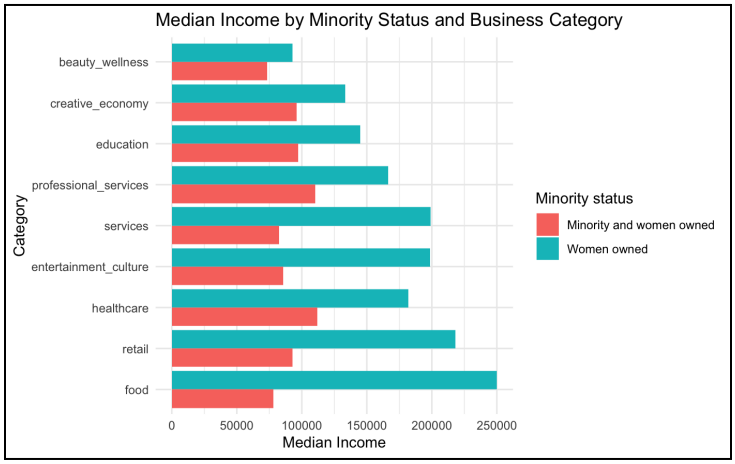



Figure 2. Median income of business types by minority status.

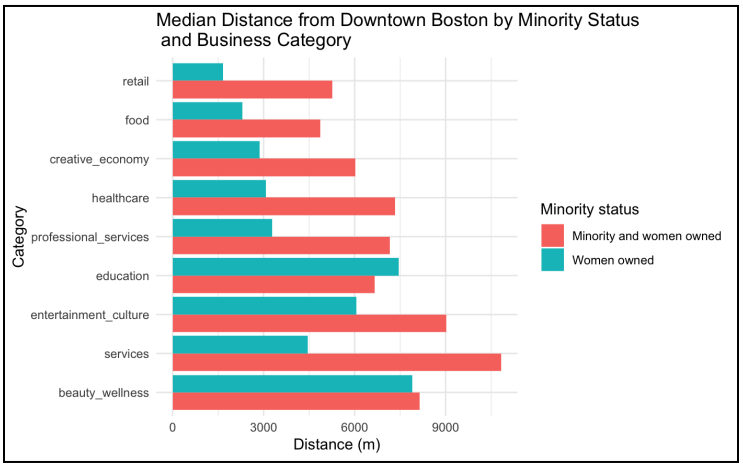



Figure 3. Median distance from downtown Boston, defined as Government Center, by business type and minority status.

These results were validated by our principal component analysis on the businesses. We mapped principal component one, which accounted for the plurality of variance in the data, in which median income and distance from Boston were strongly inversely correlated (Figures 4, 5).

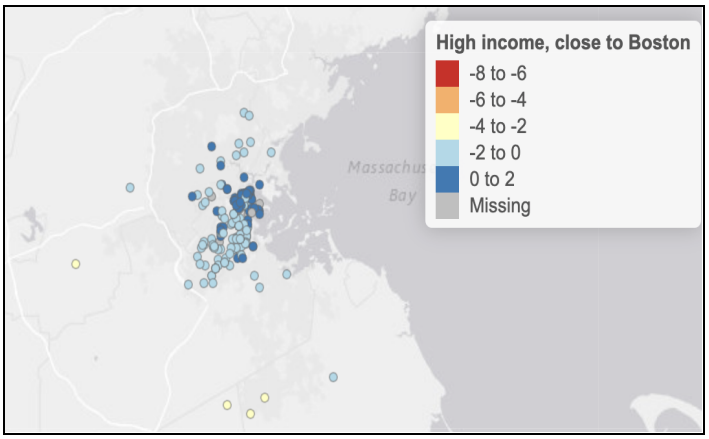



Figure 4. Map of principal component 1 of principal component analysis: businesses in high income, close to Boston areas.

```
Loadings:
                                        PC1    PC2    PC3
estimate_families_median_income_dollars  0.791          0.611
distance                                -0.792          0.611
angle                                            1.000

                 PC1   PC2   PC3
SS loadings    1.254 1.000 0.746
Proportion Var 0.418 0.333 0.249
Cumulative Var 0.418 0.751 1.000
```

Figure 5. Loadings of the first three principal components from principal component analysis.

From these analyses we found evidence that suggests there are discrepancies in the median income and distance from Boston between minority and non-minority women-owned businesses across all business types, although there is no clear relation between minority status and business type. Due to the discrepancies in these variables by minority status, we believe these variables in the data are sufficient to predict whether a business is minority-owned.

Models

We performed all these analyses in R (2023.06.1+524). In these models our outcome variable is if a women-owned business is also minority owned (1) or not (0). We use distance ($X_{dist}$), bearing ($X_{angle}$), median income of the census tract in which the business is location ($X_{income}$), and business type (nine recategorized dummy variables, abbreviated as $X_{business}$ in the models below) as predictor variables. We split all of our data into 50% train and 50% test data, if a validation set is used then the test set is split 50/50.

a. k-Nearest Neighbor (kNN)

We chose kNN because we hypothesize that businesses with similar location and income attributes will have similar demographic information. We exclude business type dummy variables from this analysis because kNN can only take continuous data.

We run kNN on k (the number of neighbors) = [1,20] to optimize for the k that gives the highest accuracy. Our equation is:

$$P_{max}(Y_{demographic} = j | X_{dist}, X_{angle}, X_{income}) = \frac{1}{K} \sum_{i \epsilon N_0} I(y_i = j)$$

Where we calculate the maximum probability of $Y = j$, where $Y$ is the outcome and $j$ is the class of minority (1) or not (0).

b. Random forest (rf)

We next employ random forests to look for predictive accuracy given binary splitting on variables. Our random forest model averages across 500 decision trees, and tests each of the twelve variables at each leaf, to get the most accurate model. Random forests calculate the Gini impurity at each node, using the following equation [7]:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Where $j$ is the node, $ni_j$ is the importance of node $j$, $w_j$ is the weighted number of samples at node j, $C_j$ is the impurity at node $j$, and $left_j$ and $right_j$ are the nodes coming from the left and right split, respectively, from node $j$ [7].

c. Multivariate binary logistic regression, forward stepwise selection

We selected multivariate binary logistic regression because the model can tell us the statistical significance of each variable as well as the magnitude of its effect (from the coefficients). We first run a multivariate binary logistic regression using all of our variables. Next, using forward stepwise selection and Akaike Information Criterion (AIC) to assess error and penalize for additional variables, we select the best combination of predictor variables that minimizes AIC. We use the following general formula:

$$p(X) = \frac{e^{\beta_0 + \beta_{dist} X_{dist} + \beta_{angle} X_{angle} + \beta_{income} X_{income} + \beta_{business} X_{business}}}{1 + e^{\beta_{dist} X_{dist} + \beta_{angle} X_{angle} + \beta_{income} X_{income} + \beta_{business} X_{business}}}$$

Where $p(X)$ follows the sigmoid function and gives us the probability of being of class 0 (non-minority) or 1 (minority). We use 0.5 as the cutoff, where any probabilities greater than or equal to 0.5 are assigned to class 1 and the rest to class 0. The β are model coefficients.

## III. Results

We ran this study with the goal of understanding the correlation between business types, location, and socioeconomic status with whether a women-owned business is also minority owned. The kNN model utilized distance, bearing, and median income for predicting minority ownership and achieved a maximum accuracy of 71.1% on the training and test data for k = 15 (Figures 6, 7). This accuracy suggests that businesses with similar location and socioeconomic status of surrounding areas have similar demographics of the owner (minority or not).

After training on the training dataset, the random forest model achieved a 60.5% accuracy on the test data set. From this model, we find that median income, distance, creative economy, and angle, respectively, have the highest importance, meaning that the model's accuracy would decrease the most if the variables were to be excluded. Business types of services, entertainment and culture, and beauty and wellness had practically no importance and therefore no impact.

Business types of professional services, food, and healthcare had negative importance, meaning the model would perform better were they to be excluded (Figure 8).

The logistic regression model using all variables achieved an accuracy of 66.5% (Figure 9). Only the median income variable was statistically significant (Figure 10). Forward stepwise selection narrowed our predictor variables down to median income and services, with a 65.4% accuracy (Figures 11, 12). Again, only median income was statistically significant.

The results from all three models indicate a notable, albeit not definitive, correlation between location, business type, and socioeconomic status with the likelihood of a business being minority-owned. The findings of this study support our hypothesis that minority-owned businesses tend to be located in areas of lower socioeconomic status. However, we did not find conclusive evidence to say that minority-owned businesses are predominantly in the service or retail industry.

The kNN model showed the strongest predictive accuracy, suggesting that businesses with similar location and income of surrounding areas have similar minority status. Median income of the area was one of the best predictors across all three models. Business types did not add predictive accuracy in random forest, most business types decreased the accuracy or had a net zero effect, and in logistic regression only the service work business type was kept during forward selection. These findings suggest the median income and distance from Boston are the best predictors of minority status of women-owned businesses. The tables and figures from all of our models are available in the Appendix.

## IV. Discussion

Our study investigated the correlation between the geographic location, business type, and socioeconomic level of the surrounding area to determine if women-owned businesses in the Greater Boston area were minority-owned. By focusing on these predictors, we aimed to provide insight to policymakers to be able to provide additional economic support to women of color business owners. This perspective is vital in a world where gender and racial inequities persist in the business landscape.

Our research employed k-Nearest Neighbor (kNN), Random Forest, and Logistic Regression models to predict whether a women-owned business is minority-owned using location, business type, and surrounding area's socioeconomic status. Our results from the three models show that location and socioeconomic level (proxied through median income of the census tract) are key in determining whether the women-owned businesses are minority-owned. The business type did not provide additional predictive accuracy.

The results highlight the importance of location and economic status in understanding minority women-owned businesses. We suggest that the government should consider the socioeconomic status of their locations when they show support to these businesses. The Policies that were created to resolve the economic imbalance in lower-income areas should also be parallel with helping the growth of minority women-owned businesses which can help to promote economic and gender equality.

This study is limited in that it only focuses on the Greater Boston area so findings may not be applicable to other major US cities such as New York City or Los Angeles, which are also famous for their multiculturalism and income disparities. Our study also only analyzes women-owned businesses, so the predictive measures may not be the same across other genders. Finally, this project viewed minority status as a binary variable – person of color or not – where in reality disaggregating race and ethnicity would likely show different results.

## V. Conclusion

Our project investigates if women-owned businesses' geographic location, business type, and socioeconomic level are predictive of if the business is also minority owned. Based on the result of our models, we find a significant relationship between socioeconomic status of the surrounding area, location, and the minority status of the business owner. We used the geographic clustering of these businesses and their level of socioeconomic status to provide insights for policymakers to provide additional resources and economic support for minority and women-owned businesses based on the location and median income.

Further research can be done by looking at patterns across other major cities in the US, and see how different economic initiatives in different cities affect business placement by race and gener. Other future research should look at gender and race intersectionalities, to look at how business type, location, and socioeconomic status vary across men and women and different races and ethnicities.

References

[1] Conway, Lou, and Alison Sheridan. “Women, Small Business and Regional Location.” *Rural Society* 15, no. 1 (2005): 55–76.

[2] Merrett, Christopher D., and John J. Gruidl. “Small Business Ownership in Illinois: The Effect of Gender and Location on Entrepreneurial Success.” *The Professional geographer* 52, no. 3 (2000): 425–436.

[3] Robb, Alicia M. “Entrepreneurial Performance by Women and Minorities: The Case of New Firms.” *Journal of Developmental Entrepreneurship* 7, no. 4 (2002): 383-.

[4] Dayanim, Suzanne Lashner. “Do Minority-Owned Businesses Face a Spatial Barrier? Measuring Neighborhood-Level Economic Activity Differences in Philadelphia: Neighborhood-Level Economic Activity Differences.” *Growth and Change* 42, no. 3 (2011): 397–419.

[5] Department of Innovation and Technology. *Women-Owned Businesses Data*. October 2023. Distributed by Boston Government. https://data.boston.gov/dataset/9772abfe-4fbe-449a-9928-f4bd4538d983/resource/91ff216b-4f54-4a2c-af00-2c0f987070ee/download/tmpv4an4j4d.csv

[6] US Census Bureau. *Median Income in the Past 12 Months (In 2021 Inflation Adjusted Dollars).* 5-Year 2021 American Community Survey. https://data.census.gov/table/ACSST5Y2021.S1903?q=median%20income&g=040XX00US25$1400000

[7] Ronaghan, Stacey. “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark” *Towards Data Science.* May 11, 2018. https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Appendix

## I. Figures

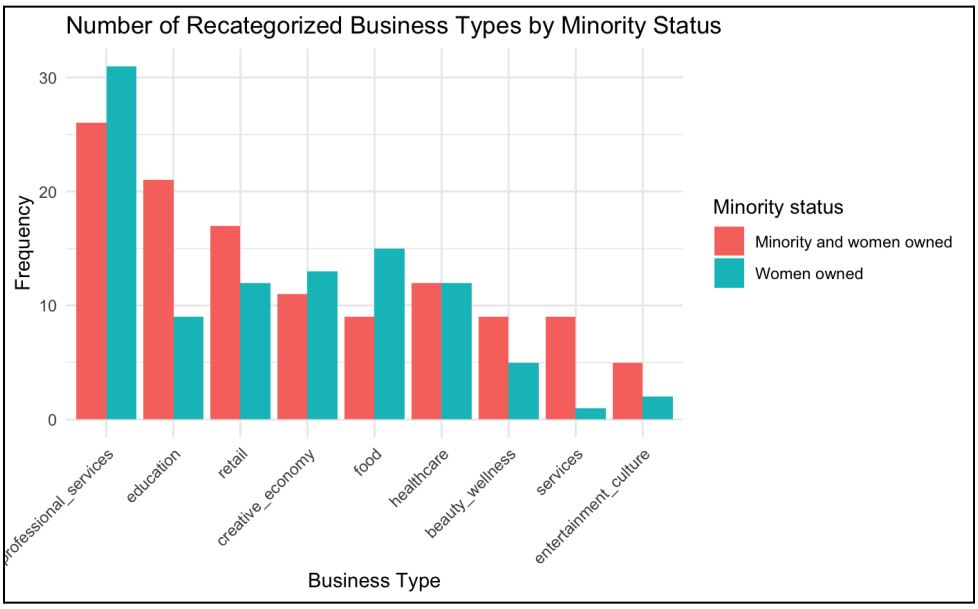



Figure 1. Number of recategorized business types by minority status.

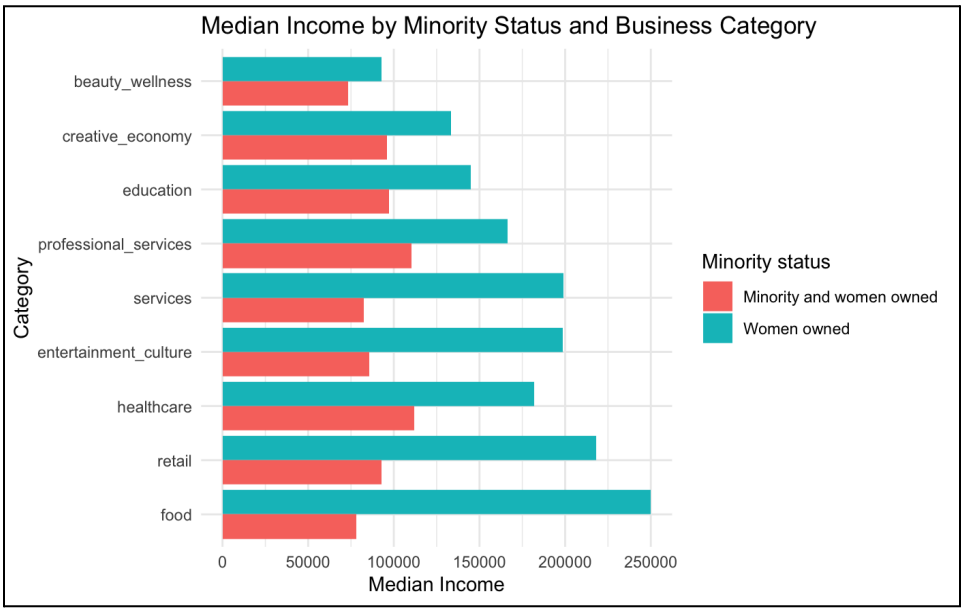



Figure 2. Median income of business types by minority status.

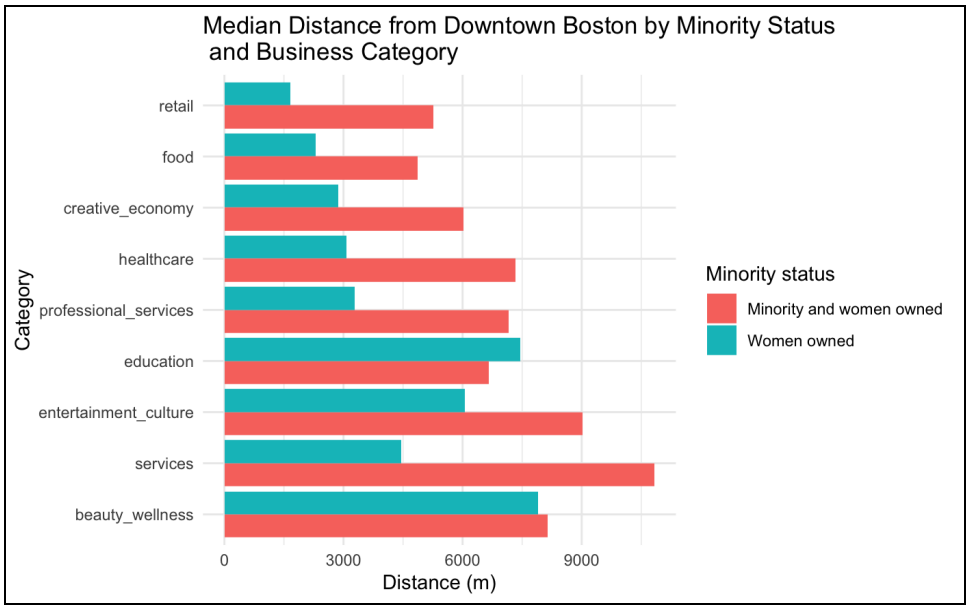



Figure 3. Median distance from downtown Boston, defined as Government Center, by business type and minority status.

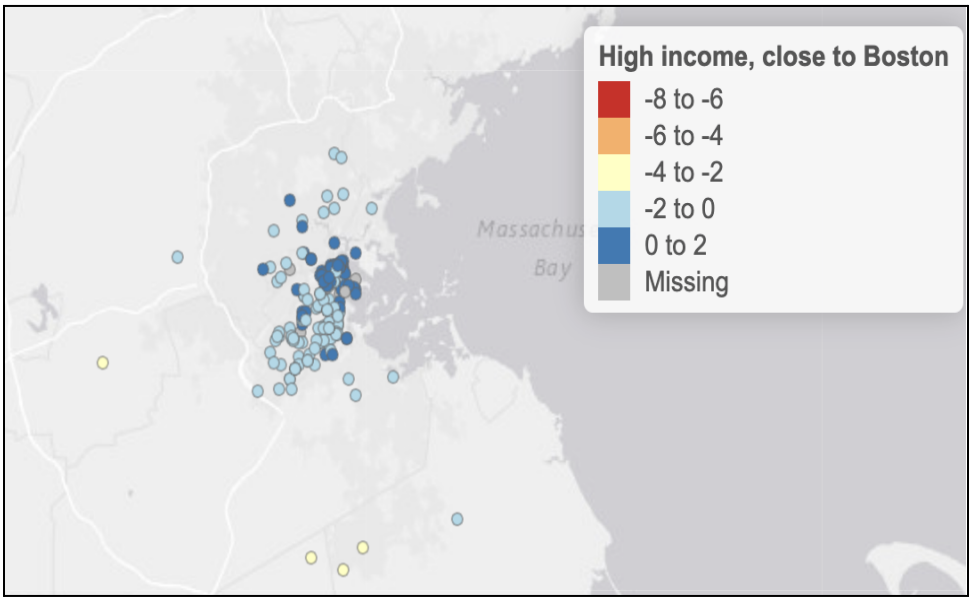



Figure 4. Map of principal component 1 of principal component analysis: businesses in high income, close to Boston areas.

```
Loadings:
                                     PC1    PC2    PC3
estimate_families_median_income_dollars  0.791          0.611
distance                                -0.792          0.611
angle                                           1.000

                 PC1   PC2   PC3
SS loadings    1.254 1.000 0.746
Proportion Var 0.418 0.333 0.249
Cumulative Var 0.418 0.751 1.000
```

Figure 5. Loadings of the first three principal components from principal component analysis.

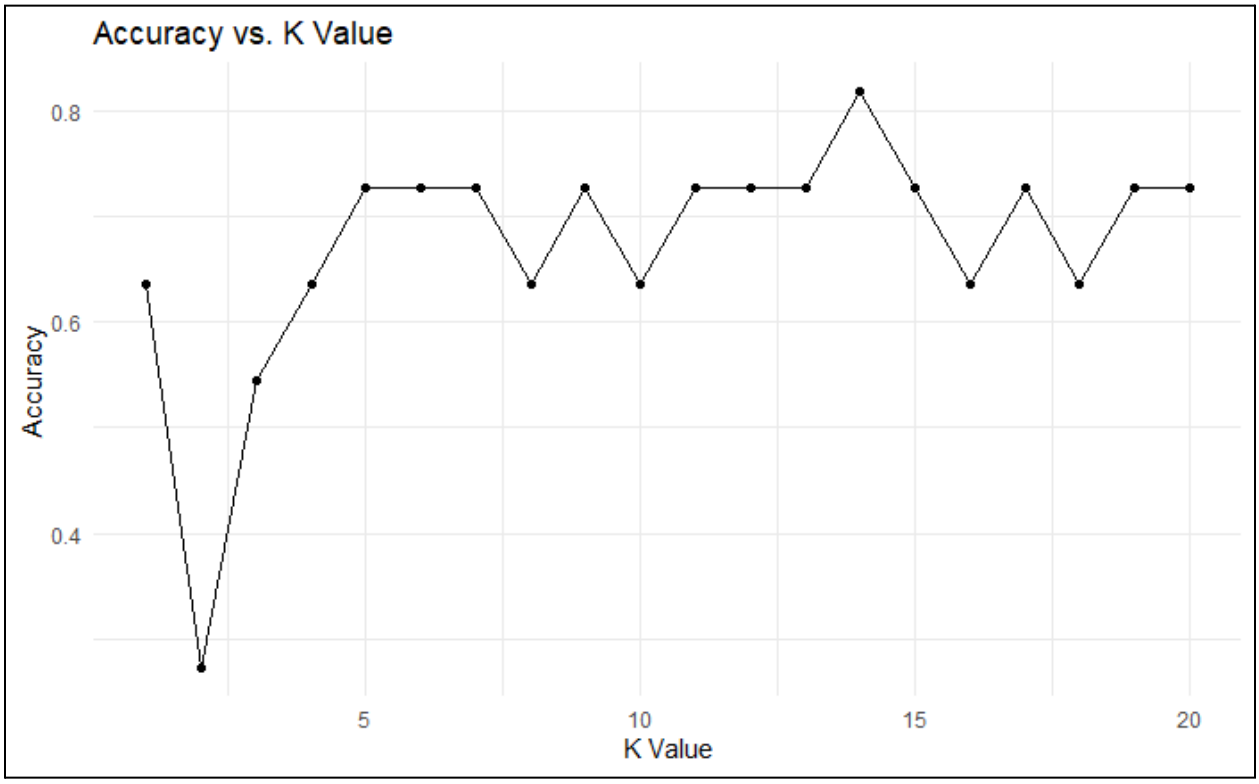



Figure 6. Accuracy vs k-value of k-Nearest Neighbors (kNN) model on training data for predicting minority status of women-owned businesses using median income, and distance and angle from Boston.

```
predicted_demographic Minority-owned N/A
       Minority-owned              17   4
       N/A                          7  10
[1] 0.7105263
```

Figure 7. Confusion matrix and accuracy for test data set of k-Nearest Neighbor model to predict minority status of women-owned businesses using median income, distance and angle from Boston.

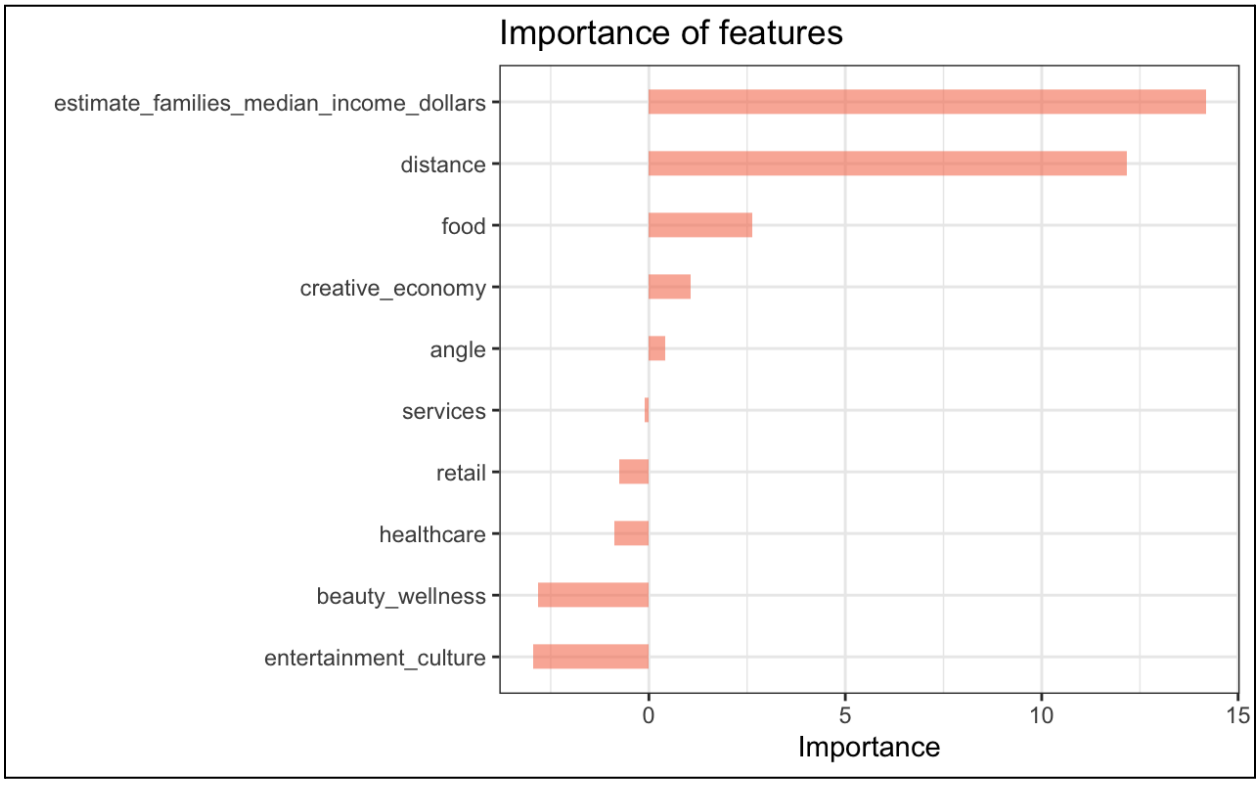



Figure 8. Feature importance of predictor variables on test data set from random forest model.

```
predicted_class  0  1
              0 46 25
              1 38 79
[1] 0.6648936
```

Figure 9. Confusion matrix and accuracy for multivariate logistic regression model using all predictor variables to predict if a women-owned business is also minority owned (1) or not (0).

```
Call:
glm(formula = minority ~ distance + angle + estimate_families_median_income_dollars +
    professional_services + entertainment_culture + beauty_wellness +
    creative_economy + retail + services + food + healthcare +
    education, family = binomial, data = women_income_knn)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.961e+00  5.361e-01   3.658 0.000255 ***
distance                              -8.499e-06  1.925e-05  -0.442 0.658785
angle                                  1.990e-03  3.057e-03   0.651 0.514992
estimate_families_median_income_dollars -1.073e-05  2.507e-06  -4.281 1.86e-05 ***
professional_services                 -4.233e-01  4.145e-01  -1.021 0.307181
entertainment_culture                  8.535e-01  9.830e-01   0.868 0.385241
beauty_wellness                       -2.978e-01  6.691e-01  -0.445 0.656289
creative_economy                      -5.246e-01  5.149e-01  -1.019 0.308268
retail                                 2.534e-01  4.853e-01   0.522 0.601560
services                               1.151e+00  1.142e+00   1.008 0.313493
food                                  -8.224e-01  5.352e-01  -1.537 0.124398
healthcare                            -2.356e-01  5.349e-01  -0.440 0.659610
education                              3.073e-01  4.753e-01   0.646 0.517993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10. Summary of multivariate logistic regression using all predictor variables to predict minority status of a business.

```
Call:
glm(formula = minority ~ estimate_families_median_income_dollars +
    services, family = binomial, data = women_income_knn)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.677e+00  3.698e-01   4.535 5.75e-06 ***
estimate_families_median_income_dollars -1.057e-05  2.313e-06  -4.571 4.86e-06 ***
services                               1.482e+00  1.103e+00   1.343    0.179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11. Summary of multivariate logistic regression forward stepwise selection to predict minority status of a business.

```
predicted_class  0  1
              0 44 25
              1 40 79
[1] 0.6542553
```

Figure 12. Confusion matrix and accuracy for multivariate logistic regression model using forward stepwise selection to predict if a women-owned business is also minority owned (1) or not (0).

**II. R Markdown Code**: Knitted document appended

# Foundations Final Project: Women-Owned Businessees

**Arlyss, Deekshith, Marco, Anushka**

**2023-12-11**

# Set up

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(sf)
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```
library(terra)
```

```
## terra 1.7.39
##
## Attaching package: 'terra'
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(tidycensus)
library(tigris)
```

```
## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.
##
## Attaching package: 'tigris'
##
## The following object is masked from 'package:terra':
##
##     blocks
```

```
library(censusxy)
library(tmap)
```

```
## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,
## which was just loaded, will retire in October 2023.
## Please refer to R-spatial evolution reports for details, especially
## https://r-spatial.org/r/2023/05/15/evolution4.html.
## It may be desirable to make the sf package available;
## package maintainers should consider adding sf to Suggests:.
## The sp package is now running under evolution status 2
##      (status 2 uses the sf package in place of rgdal)
```

```
library(flexmix)
```

```
## Loading required package: lattice
```

```
library(ggplot2)
library(geosphere)
```

# Bring in data

```
women <- read_csv("women.csv") %>%
  janitor::clean_names() %>% # clean column names so easier to work with %>%
  filter(!duplicated(business_name, physical_location_address)) %>% # filter out fully d
uplicated rows with exact same entries
  mutate(state = "Massachusetts", # add column for state for geocoding
         street_address = str_to_title(street_address), #clean names
         other_information = case_when(other_information == "Minority-owned, N/A" ~ "Min
ority-owned",
                                       other_information == "Immigrant-owned, N/A" ~ "Im
migrant-owned",
                                       .default = other_information)) # this is to stand
ardize and combine forms
```

```
## Rows: 195 Columns: 107
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (20): business_name, business_type, physical_location_address, street_ad...
## dbl (77): geoid, statefp, name_y, aland, awater, intptlat, intptlon, estimat...
## lgl (10): estimate_households_percent_allocated_family_income_in_the_past_12...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
income <- tidycensus::get_acs(geography = "tract",
                              state = "Massachusetts",
                              table = "S1901",
                              year = 2021,
                              survey = "acs5")
```

```
## Getting data from the 2017-2021 5-year ACS
```

```
## Warning: • You have not set a Census API key. Users without a key are limited to 500
## queries per day and may experience performance limitations.
## ℹ For best results, get a Census API key at
## http://api.census.gov/data/key_signup.html and then supply the key to the
## `census_api_key()` function to use it throughout your tidycensus session.
## This warning is displayed once per session.
```

```
## Loading ACS5/SUBJECT variables for 2021 from table S1901. To cache this dataset for f
aster access to ACS tables in the future, run this function with `cache_table = TRUE`. Y
ou only need to do this once per ACS dataset.
## Using the ACS Subject Tables
## Using the ACS Subject Tables
## Using the ACS Subject Tables
```

```
tract_geo <- tracts(state = "Massachusetts",
                    year = 2021)
```

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   1%
  |
  |==                                                                    |   3%
  |
  |===                                                                   |   4%
  |
  |===                                                                   |   5%
  |
  |====                                                                  |   5%
  |
  |=====                                                                 |   7%
  |
  |=======                                                               |   9%
  |
  |=======                                                               |  10%
  |
  |========                                                              |  11%
  |
  |========                                                              |  12%
  |
  |=========                                                             |  12%
  |
  |=========                                                             |  13%
  |
  |==========                                                            |  14%
  |
  |==========                                                            |  15%
  |
  |===========                                                           |  15%
  |
  |===========                                                           |  16%
  |
  |============                                                          |  17%
  |
  |============                                                          |  18%
  |
  |=============                                                         |  18%
  |
  |===============                                                       |  22%
  |
  |================                                                      |  23%
  |
  |================                                                      |  24%
  |
  |=================                                                     |  24%
  |
  |=================                                                     |  25%
  |
```

```
|==================                                                    |  26%
|
|==================                                                    |  26%
|
|==================                                                    |  27%
|
|===================                                                   |  28%
|
|===================                                                   |  28%
|
|===================                                                   |  29%
|
|====================                                                  |  30%
|
|====================                                                  |  31%
|
|====================                                                  |  32%
|
|=========================                                             |  36%
|
|=========================                                             |  37%
|
|=========================                                             |  38%
|
|==========================                                            |  38%
|
|==========================                                            |  39%
|
|===========================                                           |  40%
|
|===========================                                           |  41%
|
|============================                                          |  41%
|
|============================                                          |  42%
|
|============================                                          |  42%
|
|============================                                          |  43%
|
|============================                                          |  44%
|
|============================                                          |  44%
|
|============================                                          |  45%
|
|============================                                          |  45%
|
|============================                                          |  46%
|
|============================                                          |  47%
|
```

```
|================================                                      |  48%
|
|================================                                      |  48%
|
|================================                                      |  49%
|
|================================                                      |  49%
|
|================================                                      |  50%
|
|=================================                                     |  51%
|
|=================================                                     |  52%
|
|==================================                                    |  52%
|
|==================================                                    |  53%
|
|===================================                                   |  54%
|
|===================================                                   |  55%
|
|====================================                                  |  55%
|
|====================================                                  |  56%
|
|=====================================                                 |  57%
|
|=====================================                                 |  58%
|
|======================================                                |  58%
|
|======================================                                |  59%
|
|=======================================                               |  60%
|
|=======================================                               |  61%
|
|========================================                              |  61%
|
|========================================                              |  62%
|
|=========================================                             |  62%
|
|=========================================                             |  63%
|
|=========================================                             |  64%
|
|==========================================                            |  64%
|
|==========================================                            |  65%
|
```

```
|=============================================                         |  65%
|
|=============================================                         |  66%
|
|==============================================                        |  67%
|
|==============================================                        |  68%
|
|==============================================                        |  68%
|
|==============================================                        |  69%
|
|===============================================                       |  69%
|
|===============================================                       |  70%
|
|================================================                      |  71%
|
|================================================                      |  72%
|
|================================================                      |  72%
|
|================================================                      |  73%
|
|=================================================                     |  74%
|
|=================================================                     |  75%
|
|=================================================                     |  75%
|
|=================================================                     |  76%
|
|==================================================                    |  77%
|
|==================================================                    |  78%
|
|==================================================                    |  78%
|
|==================================================                    |  79%
|
|===================================================                   |  80%
|
|===================================================                   |  81%
|
|====================================================                  |  81%
|
|====================================================                  |  82%
|
|====================================================                  |  82%
|
|=====================================================                 |  83%
|
```

```
  |=============================================================          |  84%
  |
  |==============================================================         |  85%
  |
  |===============================================================        |  85%
  |
  |================================================================       |  86%
  |
  |=================================================================      |  87%
  |
  |==================================================================     |  88%
  |
  |===================================================================    |  88%
  |
  |====================================================================   |  89%
  |
  |=====================================================================  |  89%
  |
  |======================================================================  |  90%
  |
  |======================================================================= |  91%
  |
  |========================================================================|  92%
  |
  |=========================================================================|  92%
  |
  |=========================================================================|  93%
  |
  |=========================================================================|  94%
  |
  |=========================================================================|  95%
  |
  |=========================================================================|  95%
  |
  |=========================================================================|  96%
  |
  |=========================================================================|  97%
  |
  |=========================================================================|  98%
  |
  |=========================================================================|  98%
  |
  |=========================================================================|  99%
  |
  |=========================================================================|  99%
  |
  |=========================================================================| 100%
```

```
label_acs <- tidycensus::load_variables(year = 2021,
                                        dataset = "acs5/subject") %>%
  rename(variable = name)
```

# Combine income and geography

```
income_geo <- full_join(income, tract_geo, by = "GEOID") %>%
  left_join(label_acs) %>%
  st_as_sf()
```

```
## Joining with `by = join_by(variable)`
```

# Geocode

```
women_geocoded <-
  cxy_geocode(.data = women,
              street = "street_address",
              city = "city",
              state = "state",
              zip = "business_zipcode")
```

This gives us 4 lat/lon coordinates. We are only missing 0 street addresses.

Upon visual inspection, issues come from “Commercial Wharf” and “Faneuil Hall”, so we need to give those actual addresses from Google. Other issues include missing addresses, which we can manually enter here:

# Add in addresses

```
women$street_address[women$street_address == "Faneuil Hall Marketplace"] <- "4 South Mar
ket"
women$street_address[women$street_address == "Commercial Wharf"] <- "47 Commercial Whar
f"
```

# Geocode again

```
women_geocoded <-
  cxy_geocode(.data = women,
              street = "street_address",
              city = "city",
              state = "state",
              zip = "business_zipcode")
```

# Add in geocodes of missing ones from Google

```
women_geocoded$cxy_lat[women_geocoded$street_address == "6 Liberty Square"] <- 42.35804
women_geocoded$cxy_lon[women_geocoded$street_address == "6 Liberty Square"] <- -71.05523

women_geocoded$cxy_lat[women_geocoded$street_address == "25 Dorchester Ave"] <- 42.34926
women_geocoded$cxy_lon[women_geocoded$street_address == "25 Dorchester Ave"] <- -71.0551
6

women_geocoded$cxy_lat[women_geocoded$street_address == "51 B St"] <- 42.10534
women_geocoded$cxy_lon[women_geocoded$street_address == "51 B St"] <- -70.87581
```

# Join women dataframe with income based on geography

```
# filter out null location values
women_geocoded <- women_geocoded %>%
  filter(!is.na(cxy_lat)) # filter columns with missing geography

  #convert women_geocoded to sf object
women_geocoded_sf <- women_geocoded %>%
  st_as_sf(coords = (108:109), crs = crs(income_geo))

# combine with income data
women_income <- st_join(women_geocoded_sf, income_geo) %>%
  select(!c(moe, variable)) %>% # remove margin of error column
  pivot_wider(names_from = "label",
              values_from = "estimate") %>%
  janitor::clean_names()
```

# Categorize business types into 9 different groups for modeling

```
women_income <- women_income %>%
  mutate(professional_services = ifelse(business_type == "Professional Services" | busin
ess_type == "Real Estate Broker/Owner" | business_type == "Website Design" |business_typ
e == "Professional Services, Coach women entrepreneurs start a business"|business_type =
= "Creative Economy, Professional Services, Advertising, Marketing, Branding"|business_t
ype == "Development and Construction"|business_type == "Financial Services, Healthcare,
Professional Services"|business_type == "Financial Services, Healthcare, Professional Se
rvices, Consulting" |business_type == "Construction"|business_type == "Financial Service
s" |business_type == "Creative Economy, Professional Services, Retail, Creative Agency"|
business_type == "Professional Services, Residential and Commercial Real Estate Sales an
d Leasing"|business_type == "Retail, Interior Design & Construction Project Management"|
business_type == "Communications and Public Affairs" |business_type == "Education, Finan
cial Services"|business_type == "Professional Services, Executive Search/Recruiting/Huma
n Resources Consulting"|business_type == "Clean-tech/Green-tech, Education, Healthcare,
Professional Services, Cleaning industry"|business_type == "Creative Economy, Profession
al Services"|business_type == "Real Estate"|business_type == "Life Coaching"|business_ty
pe == "Architecture"|business_type == "Coaching"|business_type ==  "Professional Service
s, software testing and data analysis - we can work with any business, any industry"|bus
iness_type == "Technology"|business_type == "Professional Services, Real Estate Brokerag
e"|business_type == "Clean-tech/Green-tech, Professional Services"|business_type == "Pro
fessional Services, Women's Empowerment Groups virtual & in person"|business_type == "Pr
ofessional Services, Real Estate"|business_type == "Clean-tech/Green-tech, Professional
Services, ELECTRICAL AND FIRE ALARM SERVICES"|business_type == "Professional Services, B
usiness Launch & Life Alignment Coaching"|business_type == "Clean-tech/Green-tech, Manuf
acturing, Professional Services, HVAC ,Mechanical, Building Automation, Clean safe Air"|
business_type == "Financial Services, Professional Services"|business_type == "Bio-tech
& Life Sciences, Clean-tech/Green-tech, Creative Economy, Education, Financial Services,
Food and Beverage, Healthcare, Manufacturing, Professional Services, Restaurant & Cateri
ng, Retail, Technology, Tourism"|business_type == "Clean-tech/Green-tech"|business_type
== "Education, Financial Services, Professional Services"|business_type == "Education, P
rofessional Services, Self Development, Communication Skills Coaching"|business_type ==
"Education, Financial Services, Professional Services, COACHING AND MENTORING"|business_
type == "Creative Economy, Financial Services, Professional Services"|business_type ==
"Legal and Investigative Group"|business_type == "Professional Services, NOTARY PUBLIC,
counseling multi services"|business_type == "Healthcare, Professional Services, Social S
ervice", 1, 0),
         entertainment_culture = ifelse(business_type ==  "Tourism, Food, Culture, Histo
ry of Boston's Chinatown"|business_type ==  "Recreational sports and social events"|busi
ness_type ==  "Professional Services, Entertainment"|business_type ==  "Tourism" |busine
ss_type ==  "Bio-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Educatio
n, Financial Services, Food and Beverage, Healthcare, Manufacturing, Professional Servic
es, Restaurant & Catering, Retail, Technology, Tourism"|business_type ==  "E-commerce Na
tural Skin Care Brand" |business_type ==  "Provide event staff"|business_type ==  "Kid e
ntertainment"|business_type ==  "Events"|business_type ==  "Party Rental Company and Eve
nt Space", 1, 0),
         beauty_wellness = ifelse(business_type ==  "apperal ,beauty and health supplie
s"|business_type ==  "Salon"|business_type ==  "Professional Services, Hair and Makeup"|
business_type ==  "Pet Care"|business_type ==  "Makeup Artistry"|business_type ==  "Beau
ty Salon"|business_type ==  "Professional Services, Hair Salon"|business_type ==  "Welln
ess"|business_type ==  "Health and wellness"|business_type ==  "Healthcare, Holistic Wel
lness"|business_type ==  "Healthcare, Hair care"|business_type ==   "Education, Professi
onal Services, Retail, hair salon"|business_type ==  "Creative Economy, Professional Ser
```

```
vices, Beauty Services"|business_type ==  "Esthetician"|business_type ==  "Fashion/Beaut
y"|business_type ==  "Spa"|business_type == "I formulate plant based skin care" |busines
s_type == "Hair Salon", 1, 0),
        creative_economy = ifelse(business_type ==  "Creative Economy"|business_type ==
"Creative Economy, Food and Beverage, Restaurant & Catering"|business_type == "Creative
Economy, Food and Beverage, Restaurant & Catering"|business_type == "Creative Economy, P
rofessional Services, Advertising, Marketing, Branding"|business_type == "Creative Econo
my, Professional Services, Retail, Creative Agency" |business_type ==  "Retail, Interior
Design & Construction Project Management"|business_type == "Bio-tech & Life Sciences, Cr
eative Economy, Healthcare"|business_type == "Art"|business_type == "Creative Agency" |b
usiness_type == "Creative Economy, Professional Services"|business_type == "Broadcast Me
dia"|business_type == "Creative Economy, Education, Retail"|business_type == "Creative E
conomy, Contemplative + Healing Arts"|business_type == "Creative Economy, Retail"|busine
ss_type == "Interior Design Services"|business_type ==  "Food and Beverage, Retail, Flor
ist"|business_type == "Creative Economy, Professional Services, Beauty Services"|busines
s_type == "Creative Economy, Graphic Design" |business_type == "Creative Economy, Manufa
cturing"|business_type == "Creative Economy, Professional Services, Photography"|busines
s_type == "Creative Economy, Education, Professional Services"|business_type == "Creativ
e Economy, Market Research, Strategy and Design"|business_type ==  "Creative Economy, Pr
ofessional Services, Retail", 1, 0),
        retail = ifelse(business_type ==  "apperal ,beauty and health supplies"|busines
s_type == "Retail, Handmade"|business_type == "Retail" |business_type == "Food and Bever
age, Retail"|business_type == "Restaurant & Catering, Retail"|business_type == "Creative
Economy, Professional Services, Retail, Creative Agency"|business_type == "Retail, Inter
ior Design & Construction Project Management"|business_type == "Creative Economy, Educat
ion, Retail"|business_type ==  "Education, Retail" |business_type == "Creative Economy,
Retail" |business_type == "Bio-tech & Life Sciences, Clean-tech/Green-tech, Creative Eco
nomy, Education, Financial Services, Food and Beverage, Healthcare, Manufacturing, Profe
ssional Services, Restaurant & Catering, Retail, Technology, Tourism"|business_type ==
"Education, Professional Services, Retail, hair salon"|business_type == "Education, Food
and Beverage"|business_type == "Creative Economy, Professional Services, Retail", 1, 0),
        services = ifelse(business_type ==  "Cleaning"|business_type == "Clean-tech/Gre
en-tech, Cleaning Company"|business_type == "Building Services-Cleaning"|business_type =
= "Towing"|business_type ==  "Green Cleaning (Residential"|business_type == "Clean-tech/
Green-tech, Education, Healthcare, Personal maid" |business_type == "Clean-tech/Green-te
ch, Education, Healthcare, Professional Services, Cleaning industry"|business_type == "J
anitorial"|business_type == "Pet Care"|business_type ==  "Clothing alteration and dry cl
eaning"|business_type == "Construction Painting"|business_type == "Provide event staff",
1, 0),
        food = ifelse(business_type ==  "Food and Beverage"|business_type == "Restauran
t & Catering, Retail"|business_type == "Restaurant & Catering"|business_type == "Food an
d Beverage, Restaurant & Catering"|business_type == "Food and Beverage, Retail"|business
_type == "Creative Economy, Food and Beverage, Restaurant & Catering"|business_type ==
"Tourism, Food, Culture, History of Boston's Chinatown"|business_type == "Creative Econo
my, Food and Beverage, Restaurant & Catering"|business_type == "Extra Virgin Olive Oil"|
business_type == "Bio-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Edu
cation, Financial Services, Food and Beverage, Healthcare, Manufacturing, Professional S
ervices, Restaurant & Catering, Retail, Technology, Tourism"|business_type == "CPG Food
Company - Frozen Meal Bites for Kids"|business_type == "Food and Beverage, Retail, Flori
st"|business_type == "Dog Bakery", 1, 0),
        healthcare = ifelse(business_type ==  "Healthcare"|business_type == "Health & W
ellness"|business_type == "Medical spa/ Day Spa"|business_type == "Forensic Science"|bus
```

```
iness_type == "Education, Healthcare, Retail, Fitness/Wellness"|business_type == "Educat
ion, Healthcare, Professional Services, Heath & Wellness"|business_type == "Suicide Prev
ention - Military and for Spanish speakers" |business_type == "Education, Healthcare, Pr
ofessional Services"|business_type == "Financial Services, Healthcare, Professional Serv
ices"|business_type == "Financial Services, Healthcare, Professional Services, Consultin
g"|business_type ==  "Bio-tech & Life Sciences, Creative Economy, Healthcare"|business_t
ype == "Clean-tech/Green-tech, Education, Healthcare, Personal maid"|business_type == "C
lean-tech/Green-tech, Education, Healthcare, Professional Services, Cleaning industry" |
business_type == "Bio-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Edu
cation, Financial Services, Food and Beverage, Healthcare, Manufacturing, Professional S
ervices, Restaurant & Catering, Retail, Technology, Tourism"|business_type == "Health an
d wellness"|business_type == "Healthcare, Holistic Wellness" |business_type ==  "Health
and Fitness" |business_type ==  "Healthcare, Professional Services, Social Service", 1,
0),
         education = ifelse(business_type ==  "Education, Professional Services, Technol
ogy, Data science / predictive modeling"|business_type ==  "Education, Professional Serv
ices"|business_type ==  "Education, Nonprofit"|business_type ==  "Education"|business_ty
pe ==  "Education, Professional Services, Non Profit"|business_type ==  "Education, Swim
School"|business_type ==  "Education, Healthcare, Professional Services, Heath & Wellnes
s"|business_type ==  "Education, Healthcare, Retail, Fitness/Wellness"|business_type ==
"Education, Healthcare, Professional Services" |business_type ==  "Education, Financial
Services" |business_type ==  "Clean-tech/Green-tech, Education, Healthcare, Personal mai
d"|business_type ==  "Clean-tech/Green-tech, Education, Healthcare, Professional Service
s, Cleaning industry"|business_type ==  "Creative Economy, Education, Retail"|business_t
ype ==  "Education, Retail"|business_type ==  "Bio-tech & Life Sciences, Clean-tech/Gree
n-tech, Creative Economy, Education, Financial Services, Food and Beverage, Healthcare,
Manufacturing, Professional Services, Restaurant & Catering, Retail, Technology, Touris
m"|business_type ==  "Education, Financial Services, Professional Services"|business_typ
e ==  "Education, Professional Services, Self Development, Communication Skills Coachin
g" |business_type ==  "Education, Financial Services, Professional Services, COACHING AN
D MENTORING"|business_type ==  "Education, Professional Services, Retail, hair salon"|bu
siness_type ==  "Creative Economy, Education, Professional Services"|business_type ==
"Education, Food and Beverage", 1, 0))
```

# Categorize minority or women owned for modeling

```
women_income <- women_income %>%
  mutate(minority = ifelse(grepl("Minority-owned", other_information), 1, 0),
         just_women = ifelse(minority == 1, 0, 1))

# recategorize minority, veteran-owned to just minority owned (variable of interest)
women_income$other_information[women_income$other_information == "Minority-owned, Vetera
n-owned"] <- "Minority-owned"
```

```
# Join latitude and longitude columns go dataframe
women_income <- left_join(women_income, women_geocoded[c("cxy_lat", "cxy_lon", "business
_name")]) %>%
  select(!c(latitude, longitude)) %>%
  rename(latitude = cxy_lat,
         longitude = cxy_lon)
```

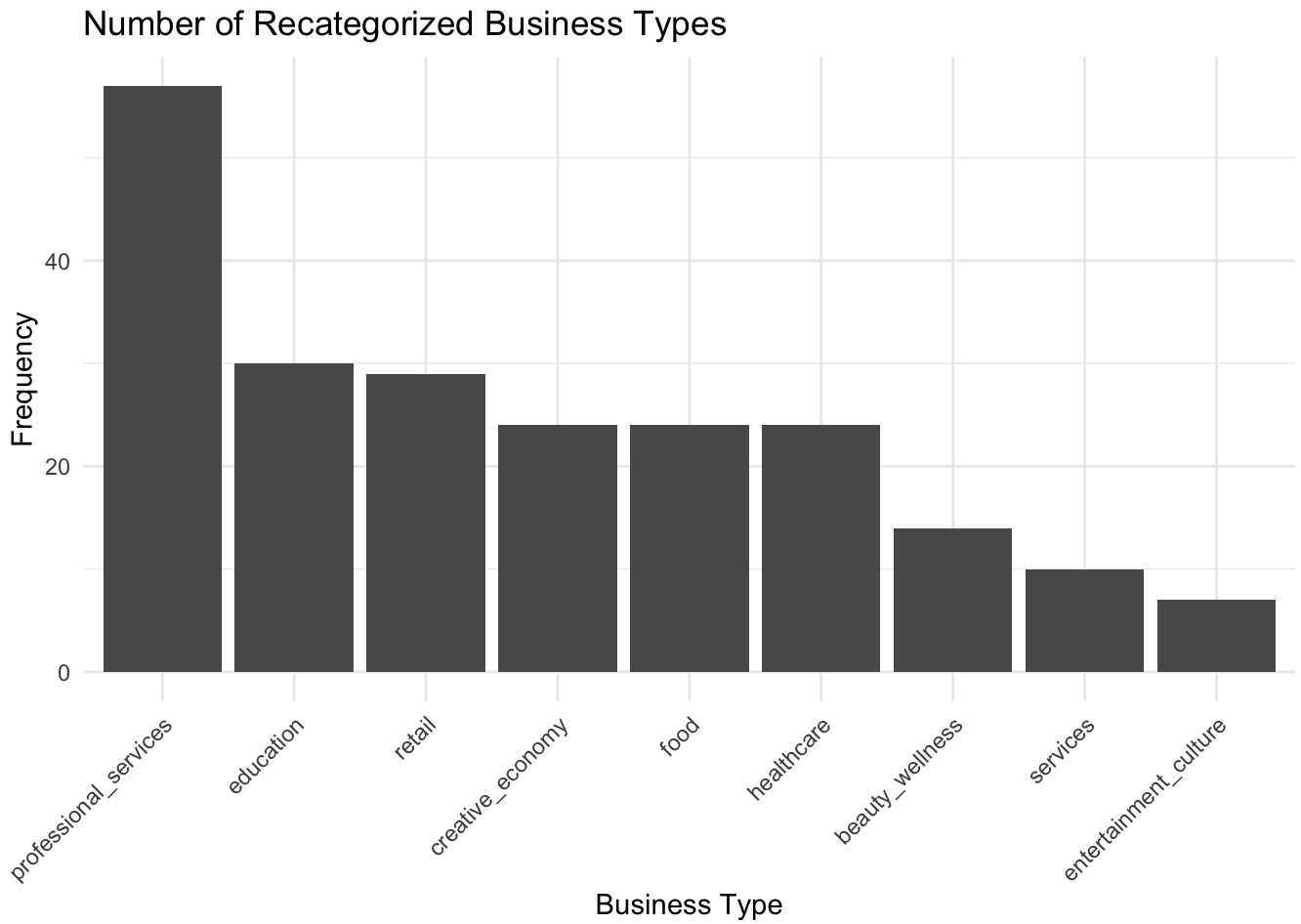
```
## Joining with `by = join_by(business_name)`
```

# Calculate distance from center of Boston and angle

```
women_income <- women_income %>%
  rowwise() %>%
  mutate(distance = distGeo(p1 = c(longitude, latitude),
                            p2 = c(-71.05908,42.36044)),
         angle = bearing(p1 = c(longitude, latitude),
                            p2 = c(-71.05908,42.36044)))
```

# Descriptive statistics

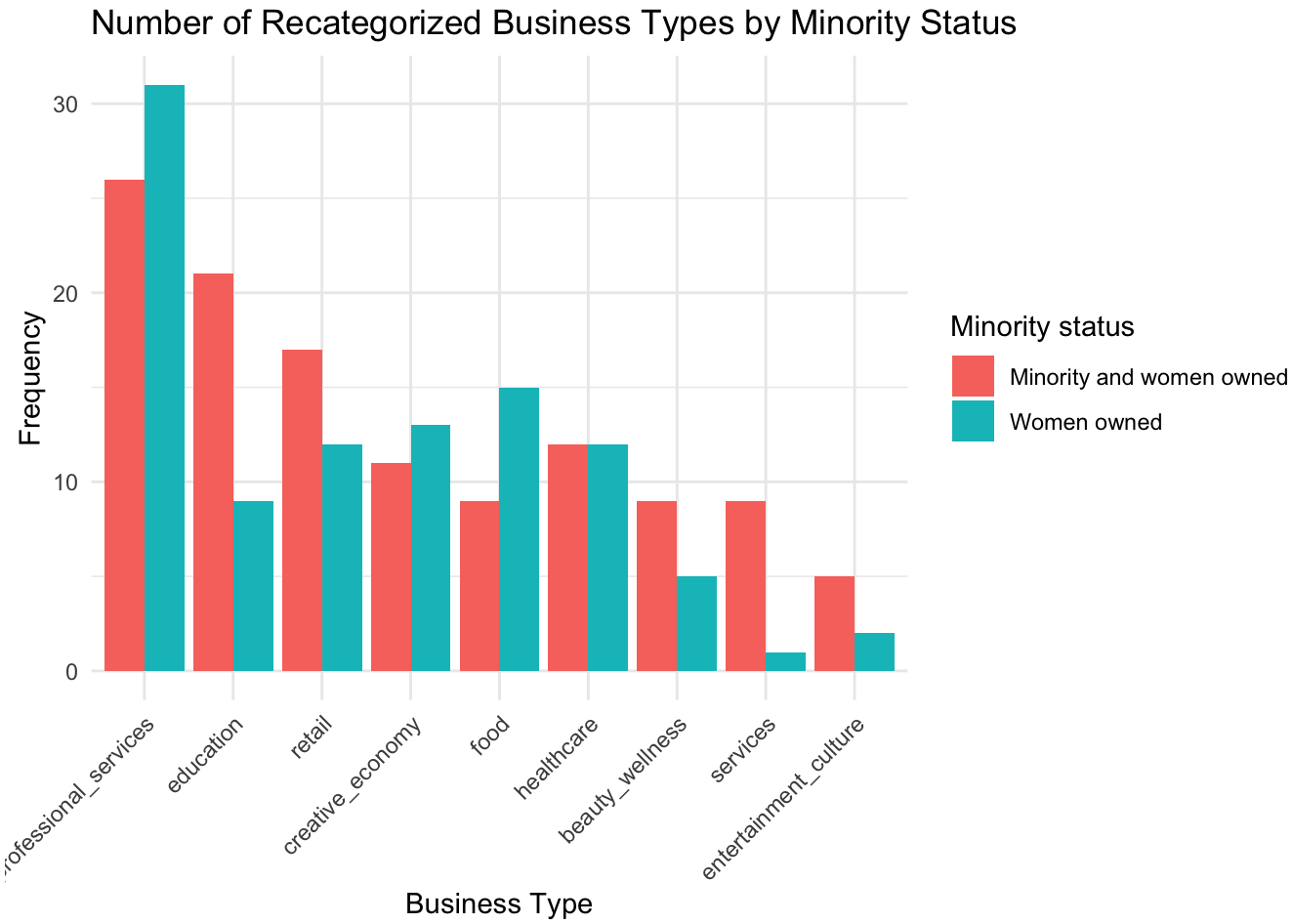
```
# Identify the most common recategorized business types
top_category_types <- women_income %>%
  st_drop_geometry() %>%
  count(professional_services, education, food, services, retail, creative_economy, ente
rtainment_culture, beauty_wellness, healthcare) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # filter for only where the business is of the category
  filter(value == 1) %>%
  group_by(name) %>%
  summarize(n = sum(n))

# Visualize the results
ggplot(top_category_types, aes(x = reorder(name, -n), y = n)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Recategorized Business Types",
       x = "Business Type",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_text(""))
```



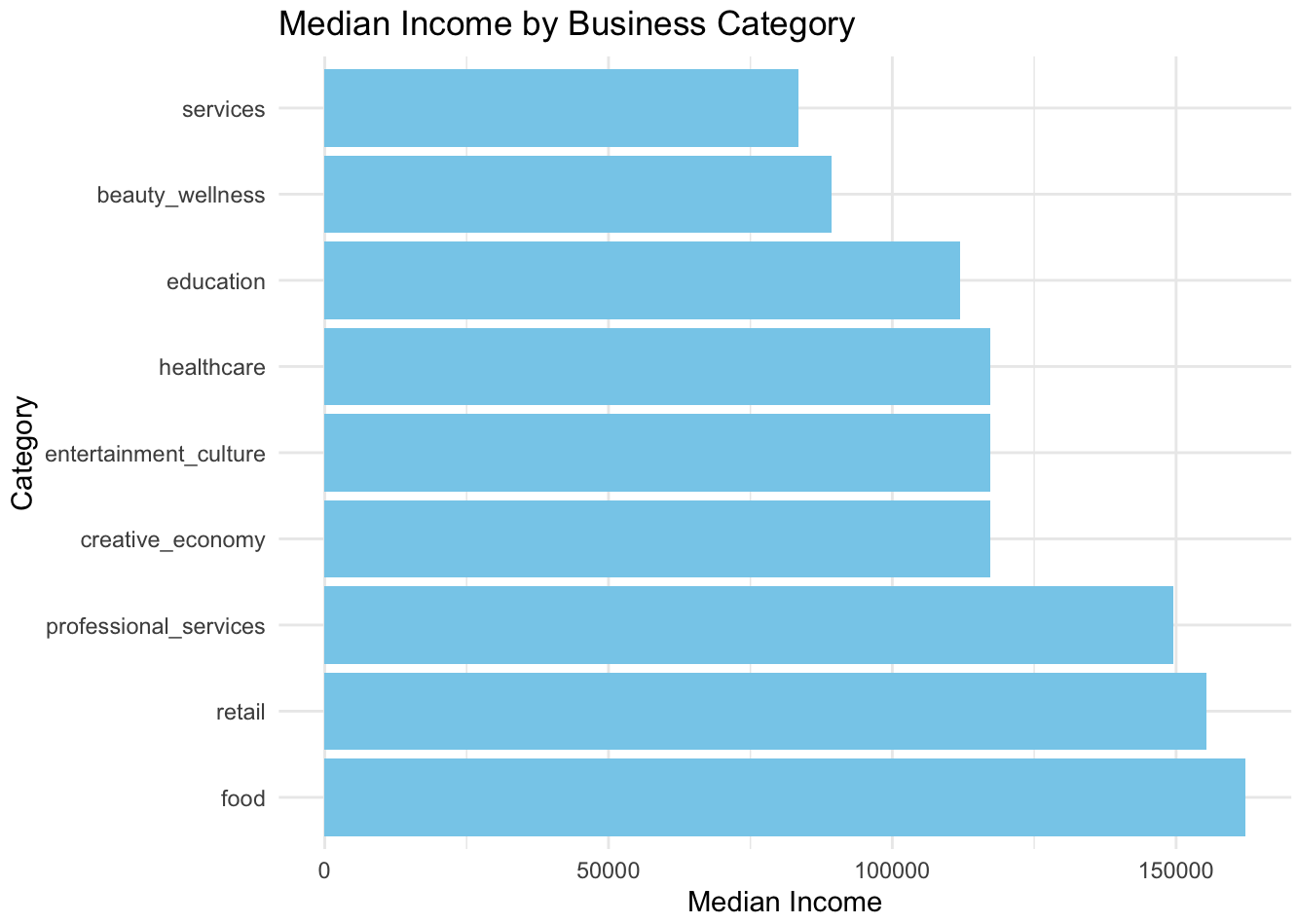

```
# Identify the most common recategorized business types
top_category_types_minority <- women_income %>%
  st_drop_geometry() %>%
  count(professional_services, education, food, services, retail, creative_economy, ente
rtainment_culture, beauty_wellness, healthcare, minority) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # filter for only where the business is of the category
  filter(value == 1) %>%
  group_by(name, minority) %>%
  summarize(n = sum(n)) %>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority and women owned", "Women ow
ned"))
```

```
## `summarise()` has grouped output by 'name'. You can override using the
## `.groups` argument.
```

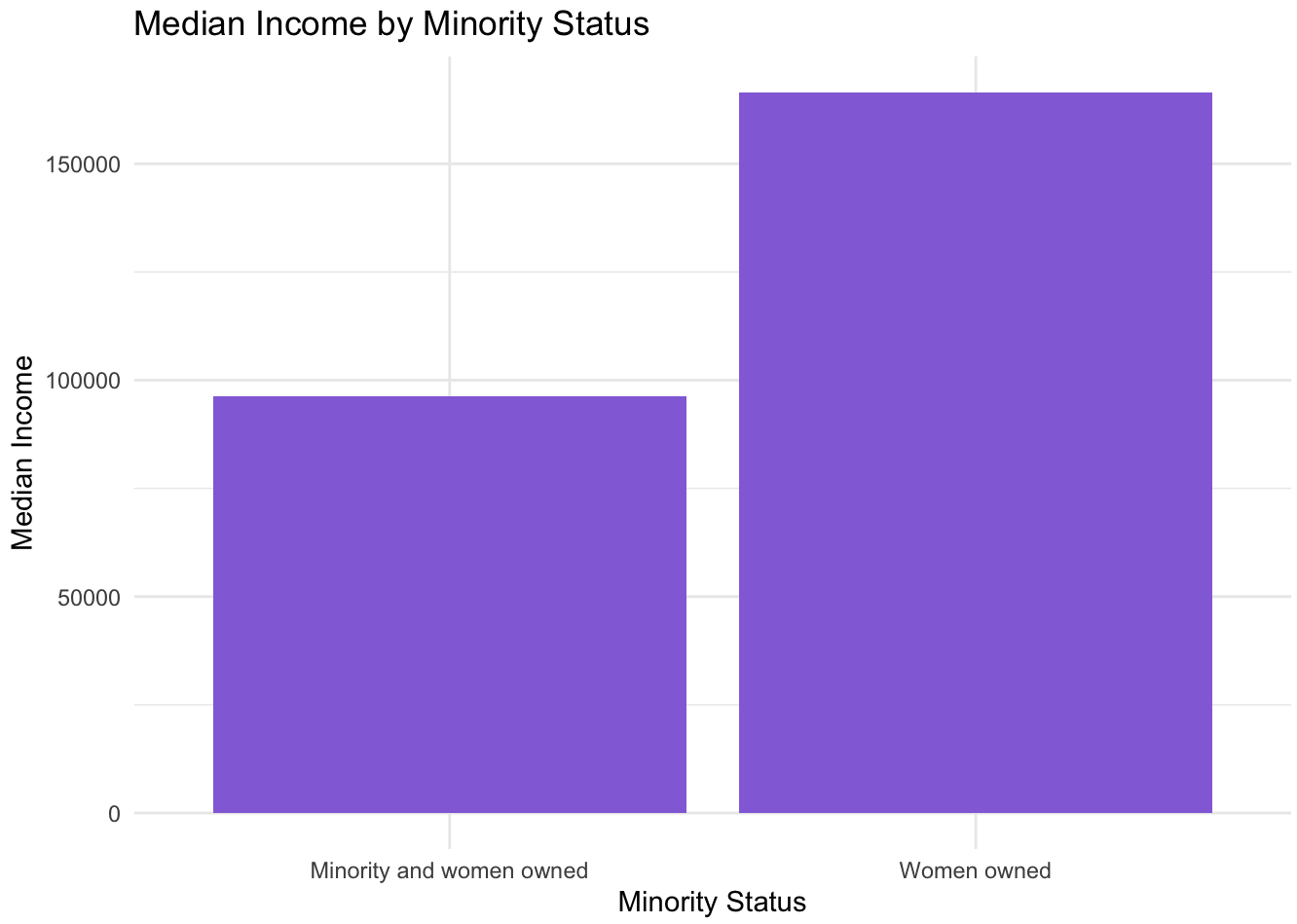
```
# Visualize the results
ggplot(top_category_types_minority, aes(x = reorder(name, -n), y = n, fill = `Minority s
tatus`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Recategorized Business Types by Minority Status",
       x = "Business Type",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.title = element_text(""))
```



```
# Descriptive statistics by category
category_stats <- women_income %>%
  st_drop_geometry() %>%
  select(professional_services, education, food, services, retail, creative_economy, ent
ertainment_culture, beauty_wellness, healthcare, estimate_families_median_income_dollar
s) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # filter for only where the business is of the category
  filter(value == 1) %>%
  group_by(name) %>%
  summarize(median_income =  median(estimate_families_median_income_dollars, na.rm = T))

# Visualize median income by category
ggplot(category_stats, aes(x = reorder(name, -median_income), y = median_income)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  theme_minimal() +
  coord_flip() +
  labs(title = "Median Income by Business Category", x = "Category", y = "Median Incom
e")
```



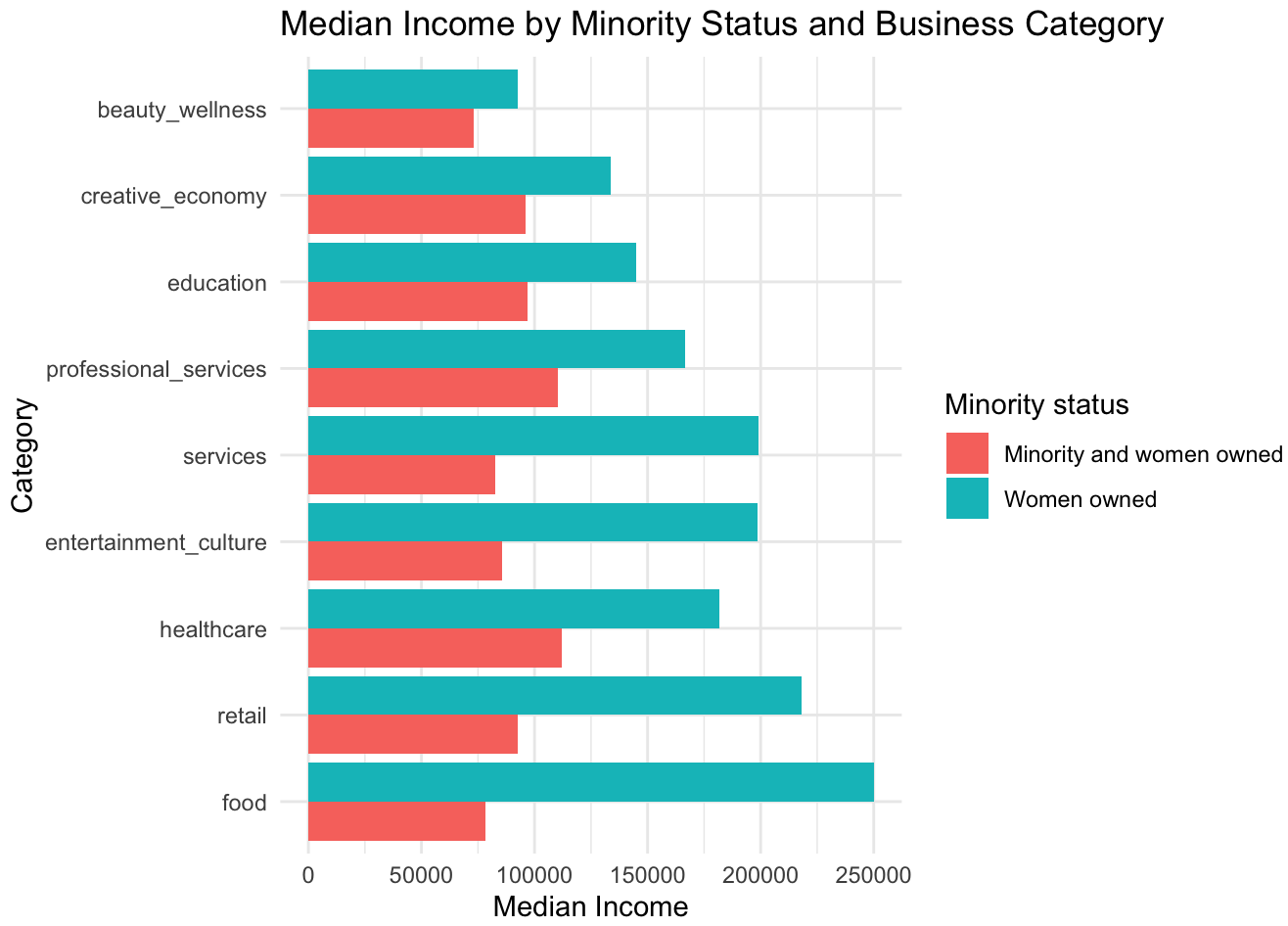
```
# Group data by minority/immigrant status and calculate median income
median_income_by_minority <- women_income %>%
  group_by(minority) %>%
  summarise(median_income = median(estimate_families_median_income_dollars, na.rm = TRU
E)) %>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority and women owned", "Women ow
ned"))

# Visualize the results
ggplot(median_income_by_minority, aes(x = as.factor(`Minority status`), y = median_incom
e)) +
  geom_bar(stat = "identity", fill = "mediumpurple") +
  labs(title = "Median Income by Minority Status",
       x = "Minority Status",
       y = "Median Income") +
  theme_minimal()
```



```
# Descriptive statistics by minority status and category
minority_category_stats <- women_income %>%
  st_drop_geometry() %>%
  select(professional_services, education, food, services, retail, creative_economy, ent
ertainment_culture, beauty_wellness, healthcare, estimate_families_median_income_dollar
s, minority) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # filter for only where the business is of the category
  filter(value == 1) %>%
  group_by(minority, name) %>%
  summarise(count = n(),
            median_income = median(estimate_families_median_income_dollars, na.rm = TRU
E)) %>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority and women owned", "Women ow
ned"))
```

```
## `summarise()` has grouped output by 'minority'. You can override using the
## `.groups` argument.
```

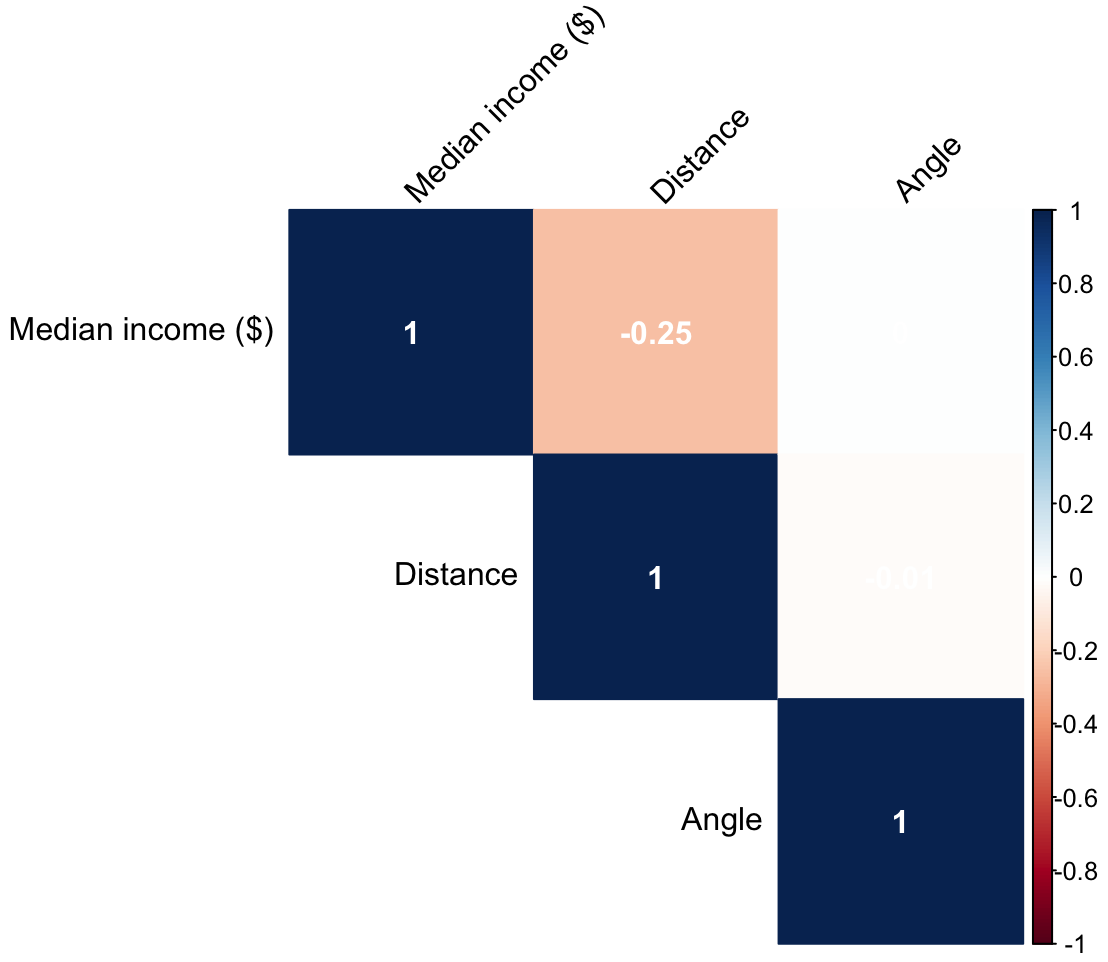
```
# Visualize median income by minority status and category
ggplot(minority_category_stats, aes(x = reorder(name, -median_income), y = median_incom
e, fill = `Minority status`)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  coord_flip() +
  labs(title = "Median Income by Minority Status and Business Category", x = "Category",
y = "Median Income")
```




```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

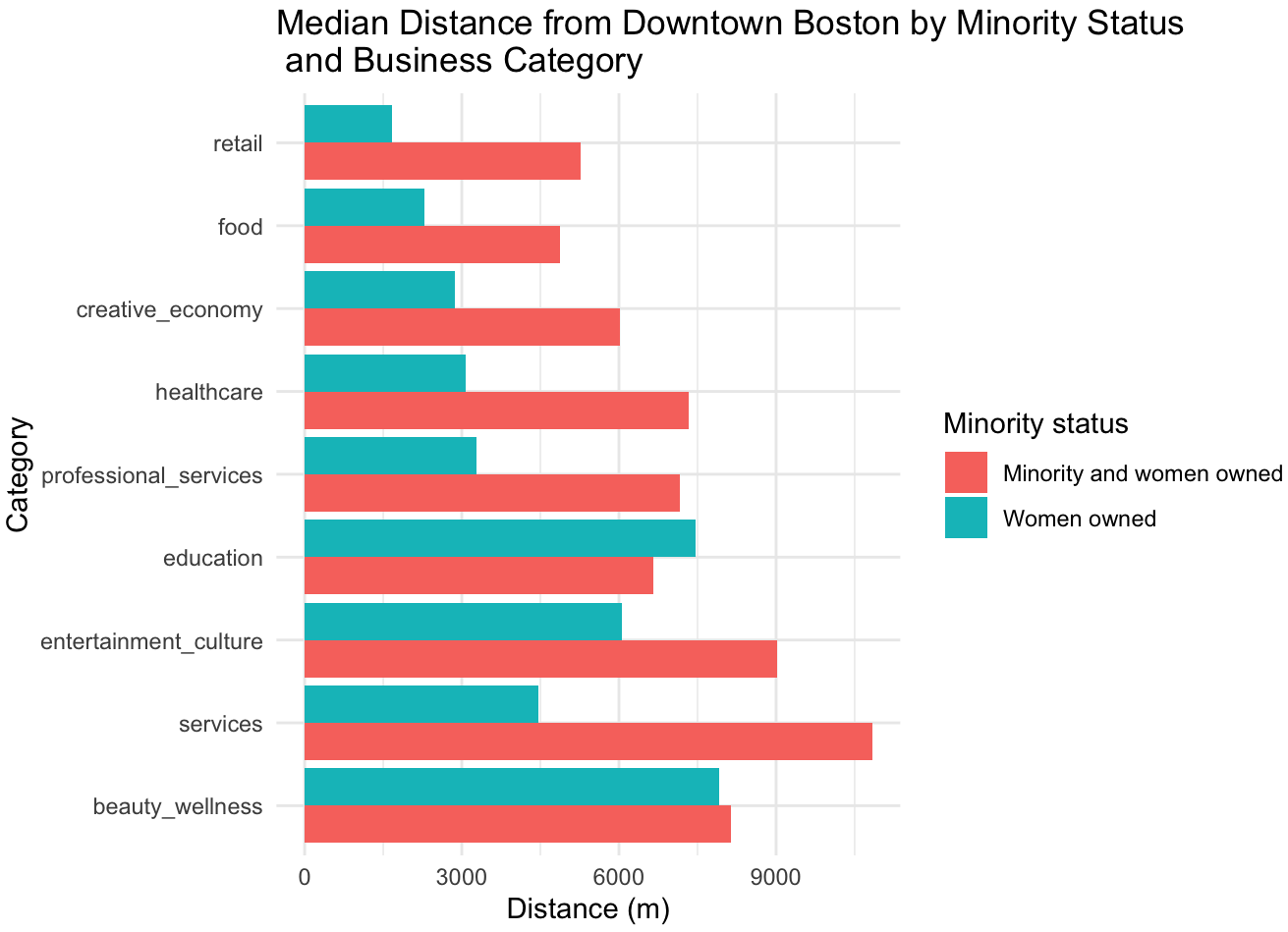
```
women_income_corrplot <- women_income %>%
  rename(`Median income ($)` = estimate_families_median_income_dollars,
         Distance = distance,
         Angle = angle)

# Correlation matrix for selected variables
correlation_matrix <- cor(women_income_corrplot[,c("Median income ($)", "Distance", "Ang
le")] %>%
  st_drop_geometry(), use = "complete.obs")


# Visualize the correlation matrix
corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt
= 45, addCoef.col = "white")
```




```
# Descriptive statistics by minority status and category

distance_category_stats <- women_income %>%
  st_drop_geometry() %>%
  select(professional_services, education, food, services, retail, creative_economy, ent
ertainment_culture, beauty_wellness, healthcare, distance, minority) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # filter for only where the business is of the category
  filter(value == 1) %>%
  group_by(minority, name) %>%
  summarise(count = n(),
            median_distance = median(distance, na.rm = TRUE)) %>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority and women owned", "Women ow
ned"))
```

```
## `summarise()` has grouped output by 'minority'. You can override using the
## `.groups` argument.
```

```
# Visualize median income by minority status and category
ggplot(distance_category_stats, aes(x = reorder(name, -median_distance), y = median_dist
ance, fill = `Minority status`)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  coord_flip() +
  labs(title = "Median Distance from Downtown Boston by Minority Status \n and Business
Category", x = "Category", y = "Distance (m)")
```

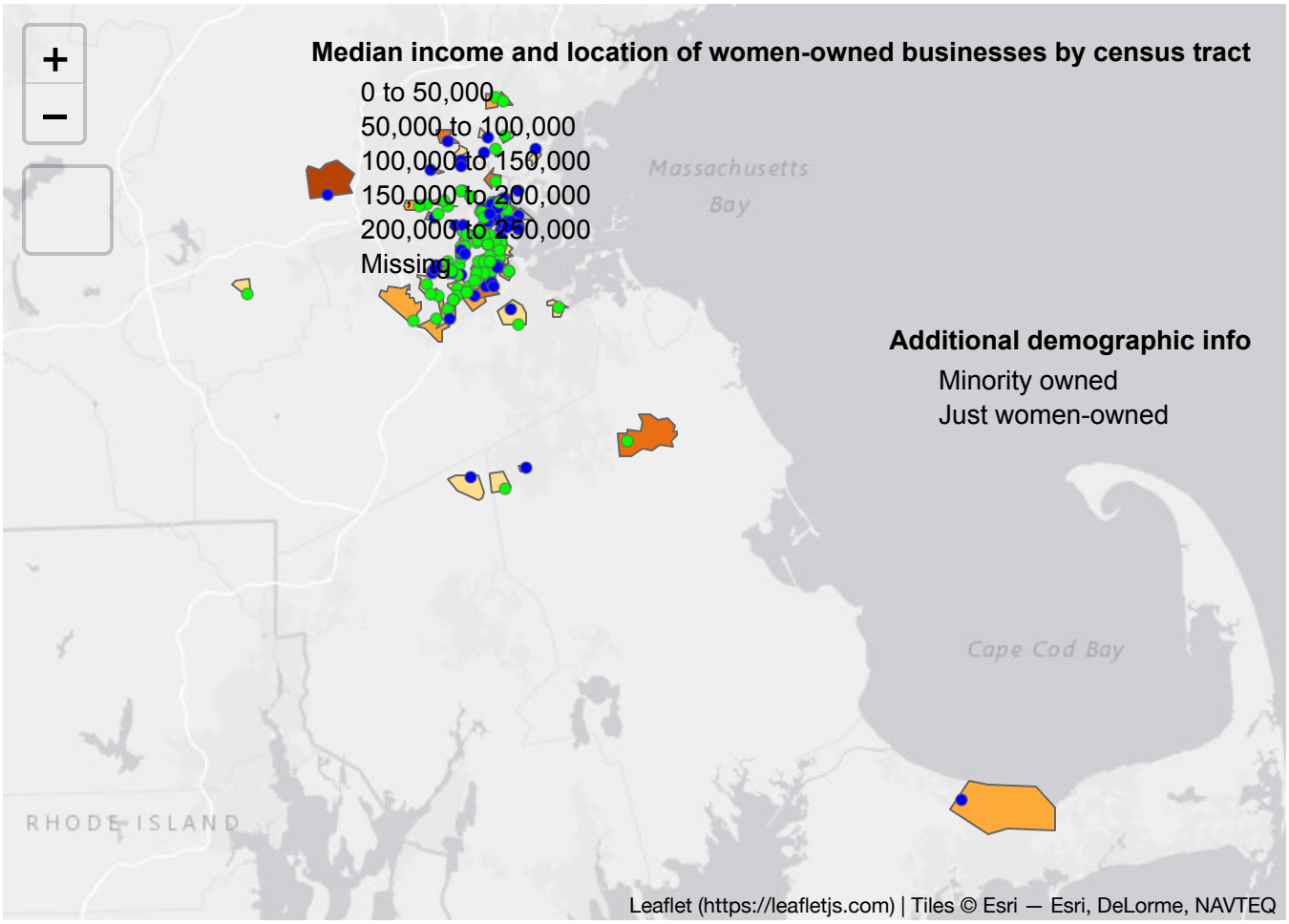



# Map of income and location of businesses

```
# create dataframe on tracts which contain women owned businesses in the dataset
income_women_tract <- st_join(income_geo, women_geocoded_sf) %>%
  filter(!is.na(business_name)
         & variable == "S1901_C01_012") %>%
  filter(!duplicated(GEOID))

# plot
map0 <- tm_shape(income_women_tract) +tm_fill(col="estimate", title="Median income and l
ocation of women-owned businesses by census tract")+tm_borders() + tmap_mode("view") +
  tm_shape(women_income) +
  tm_dots("just_women", title = "Additional demographic info", breaks = c(0,.9,1.1), lab
els = c("Minority owned", "Just women-owned"), palette=c('green','blue'))
```

```
## tmap mode set to interactive viewing
```

```
map0
```




# PCA of median income, distance, and angle

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:terra':
##
##     describe, distance, rescale
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
#subset to columns
women_income_subset <- women_income %>%
  st_drop_geometry() %>%
  select(c("estimate_families_median_income_dollars", "distance", "angle"))

# conduct pca
pca <- principal(women_income_subset, rotate="none", nfactors=3, scores=TRUE)

#show the eigenvalues
pca$values
```

```
## [1] 1.2539279 0.9999233 0.7461487
```

```
#communality closer to 1 means variable is better explained by the components
pca$communality
```

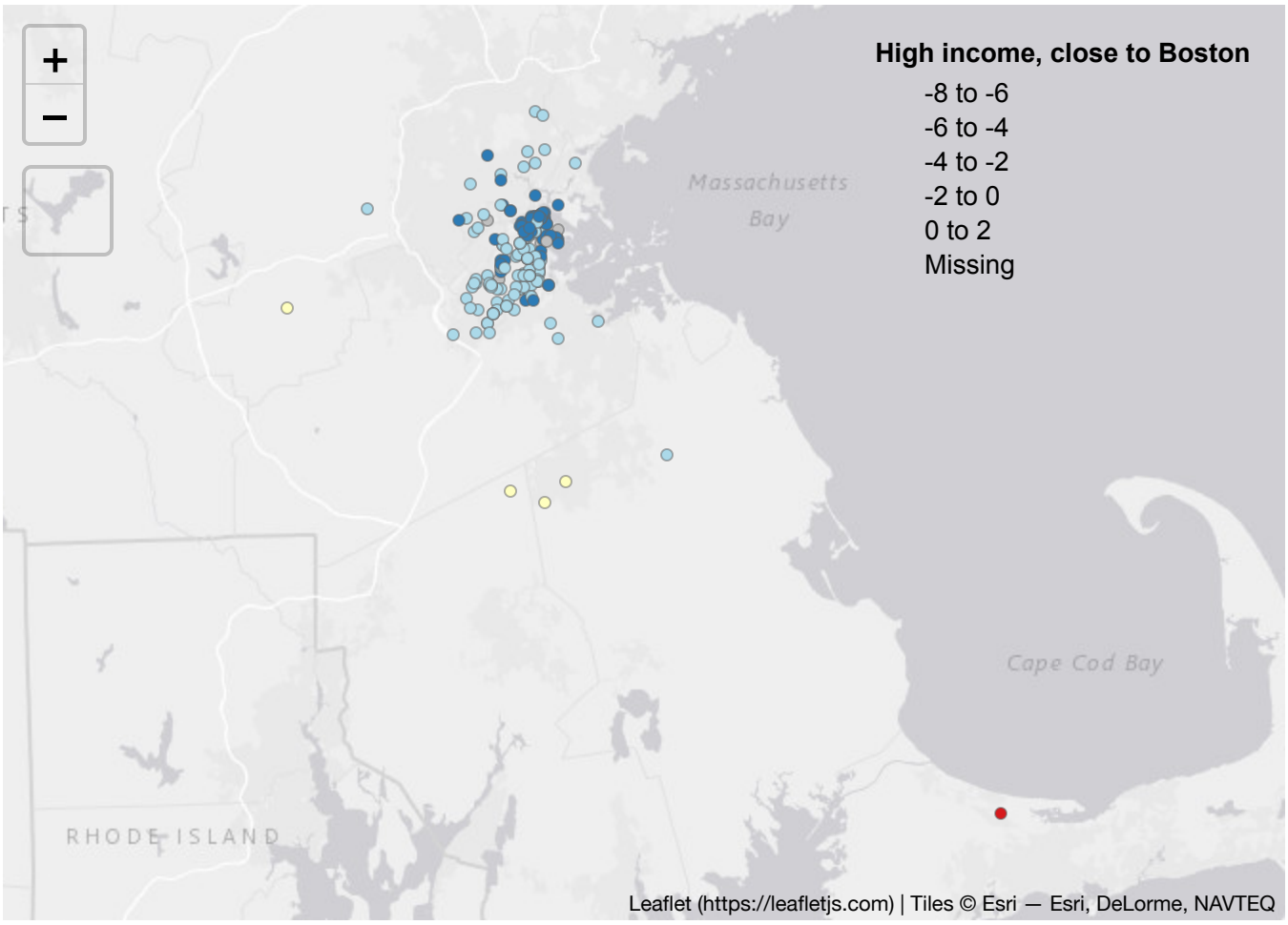
```
## estimate_families_median_income_dollars                                 distance
##                                       1                                        1
##                                   angle
##                                       1
```

```
# look at correlations
pca$loadings
```

```
##
## Loadings:
##                                         PC1    PC2    PC3
## estimate_families_median_income_dollars  0.791         0.611
## distance                                -0.792         0.611
## angle                                           1.000
##
##                  PC1   PC2   PC3
## SS loadings    1.254 1.000 0.746
## Proportion Var 0.418 0.333 0.249
## Cumulative Var 0.418 0.751 1.000
```

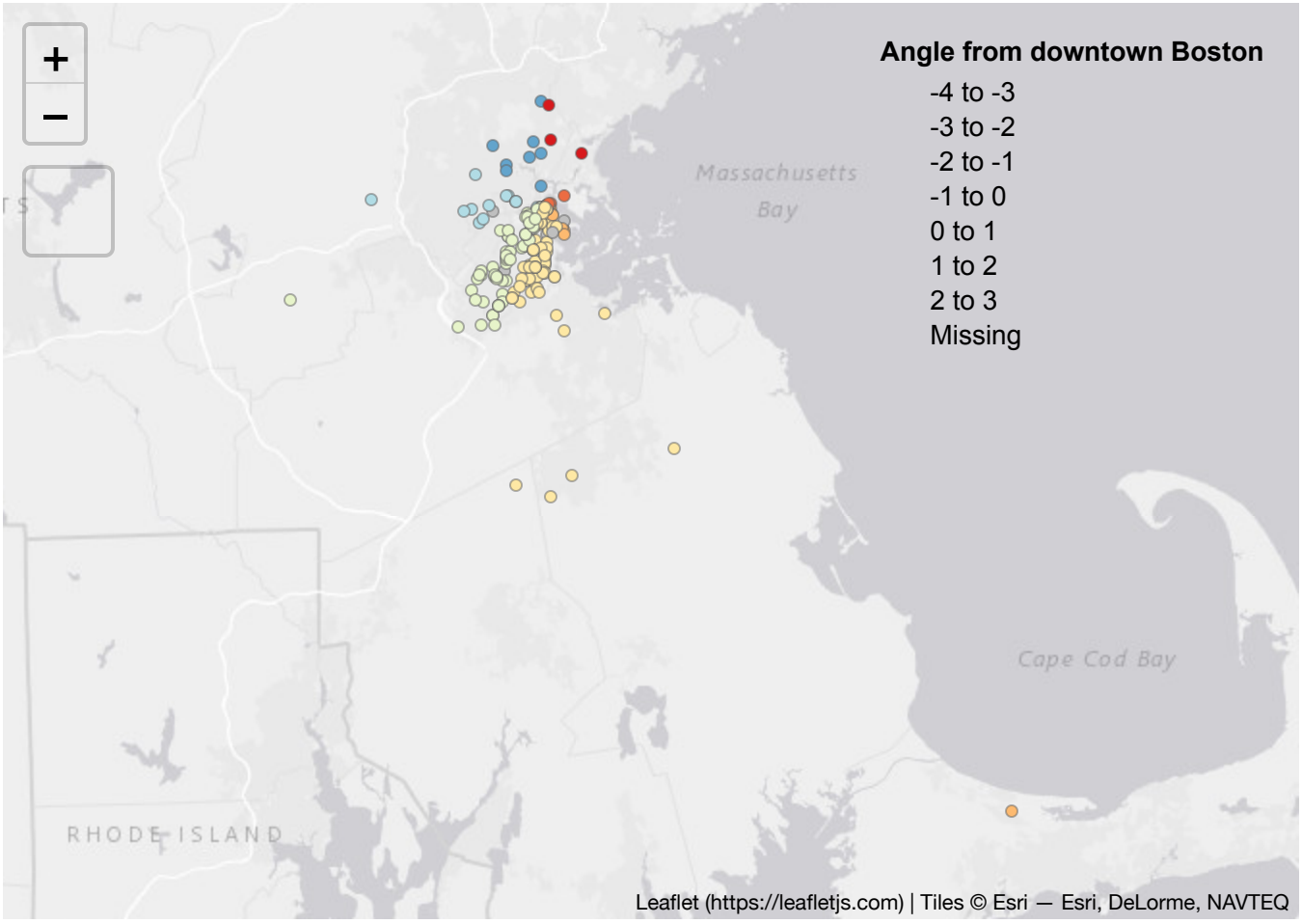
```
library(RColorBrewer)

#save the scores for each location
women_income_pca <- cbind(women_income, pca$scores)

#create palette
pc_palette <- brewer.pal(5, "RdYlBu")

#Mapping the first three components
map_pc1 <- tm_shape(women_income_pca) +tm_dots(col = "PC1", title = "High income, close
to Boston", palette = pc_palette)
map_pc1
```

```
## Variable(s) "PC1" contains positive and negative values, so midpoint is set to 0. Set
midpoint = NA to show the full spectrum of the color palette.
```



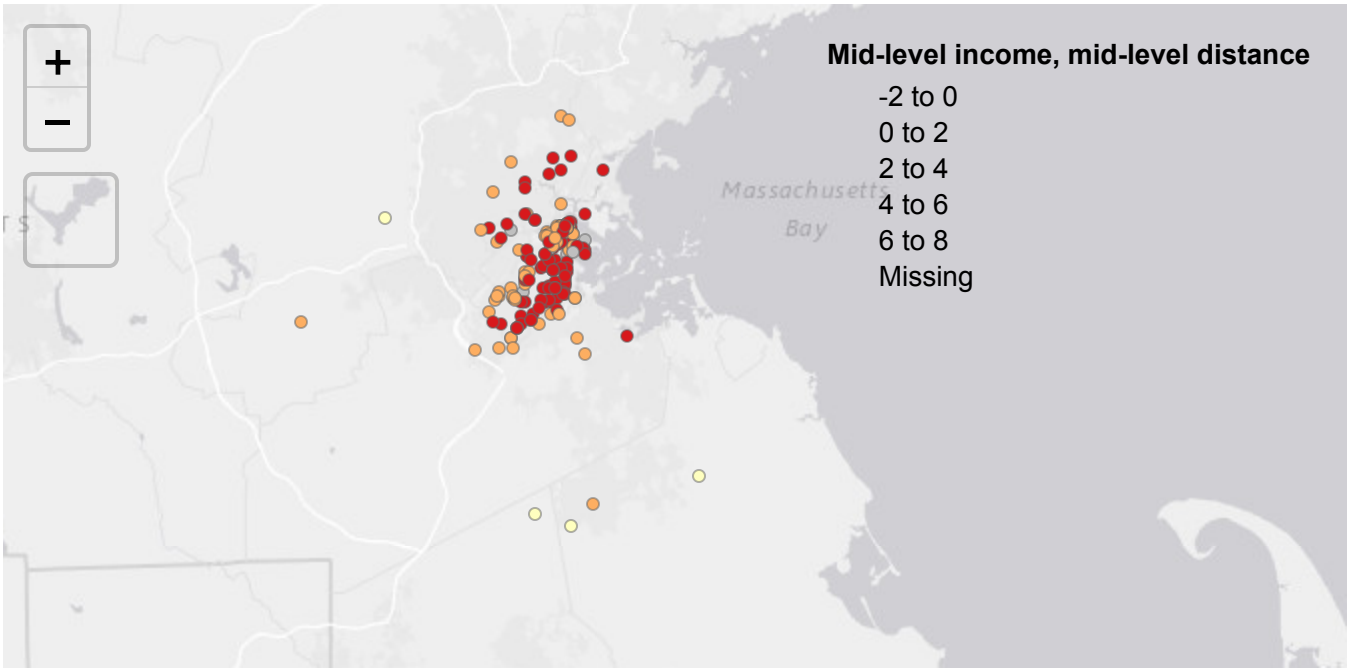

```
map_pc2 <- tm_shape(women_income_pca) +tm_dots(col="PC2", title="Angle from downtown Bos
ton", palette = pc_palette)
map_pc2
```

```
## Variable(s) "PC2" contains positive and negative values, so midpoint is set to 0. Set
midpoint = NA to show the full spectrum of the color palette.
```




```
map_pc3 <- tm_shape(women_income_pca) +tm_dots(col="PC3", title="Mid-level income, mid-level distance", palette = pc_palette)
map_pc3
```

```
## Variable(s) "PC3" contains positive and negative values, so midpoint is set to 0. Set midpoint = NA to show the full spectrum of the color palette.
```




# k Nearest Neighbor model

```
# get info on if minority or not
# filter out census tract with no residents, so no median income
# remove spatial aspect of df to get rid of geometry columns
 women_income_knn <- women_income %>%
  mutate(other_information = ifelse(grepl("Minority", other_information), "Minority-owne
d", "N/A")) %>%
  filter(!is.na(estimate_families_median_income_dollars)) %>%
  st_drop_geometry()
```

```
# Normalize the predictors individually
women_income_knn$distance_norm <- scale(women_income_knn$distance)
women_income_knn$angle_norm <- scale(women_income_knn$angle)
women_income_knn$median_norm <- scale(women_income_knn$estimate_families_median_income_d
ollars)
```

```
# Set seed for reproducibility
set.seed(12)

# Select only the normalized columns and "business_outcome" for classification
women_income_subset <- women_income_knn[, c("distance_norm", "angle_norm","median_nor
m","other_information")]

# Split data into 60% training and 40% temporary from the total number of rows
train_set_indices <- sample(1:nrow(women_income_subset), 0.6 * nrow(women_income_subse
t), replace = FALSE)
train_data <- women_income_subset[train_set_indices, ]
temp_data <- women_income_subset[-train_set_indices, ]

# Split temp_data by 50% to get 20% validation and test data each
test_set_indices <- sample(1:nrow(temp_data), 0.5 * nrow(temp_data), replace = FALSE)
test_data <- temp_data[test_set_indices, ]
validation_data <- temp_data[-test_set_indices, ]

# Print the subset of data
print(women_income_subset)
```

```
## # A tibble: 188 × 4
## # Rowwise:
##    distance_norm[,1] angle_norm[,1] median_norm[,1] other_information
##                <dbl>          <dbl>           <dbl> <chr>
##  1           -0.234         -0.212          -1.22  Minority-owned
##  2            0.592          2.85            0.212 Minority-owned
##  3           -0.0259        -0.389          -1.18  N/A
##  4           -0.620          0.601           1.25  Minority-owned
##  5           -0.554          2.68            0.945 Minority-owned
##  6           -0.566          0.664           1.53  N/A
##  7           -0.124         -0.407          -0.793 Minority-owned
##  8            0.352          0.230          -0.335 Minority-owned
##  9           -0.405         -1.08            0.118 N/A
## 10           -0.717         -0.485           1.53  Minority-owned
## # ℹ 178 more rows
```

```
# Check dimensions of the train, test, and validation data
dim(women_income_subset)
```

```
## [1] 188   4
```

```
dim(train_data)
```

```
## [1] 112   4
```

```
dim(test_data)
```

```
## [1] 38  4
```

```
dim(validation_data)
```

```
## [1] 38  4
```

```
#check for null values in train, test, validation data
any(is.na(train_data))
```

```
## [1] FALSE
```

```
any(is.na(test_data))
```

```
## [1] FALSE
```

```
any(is.na(validation_data))
```

```
## [1] FALSE
```

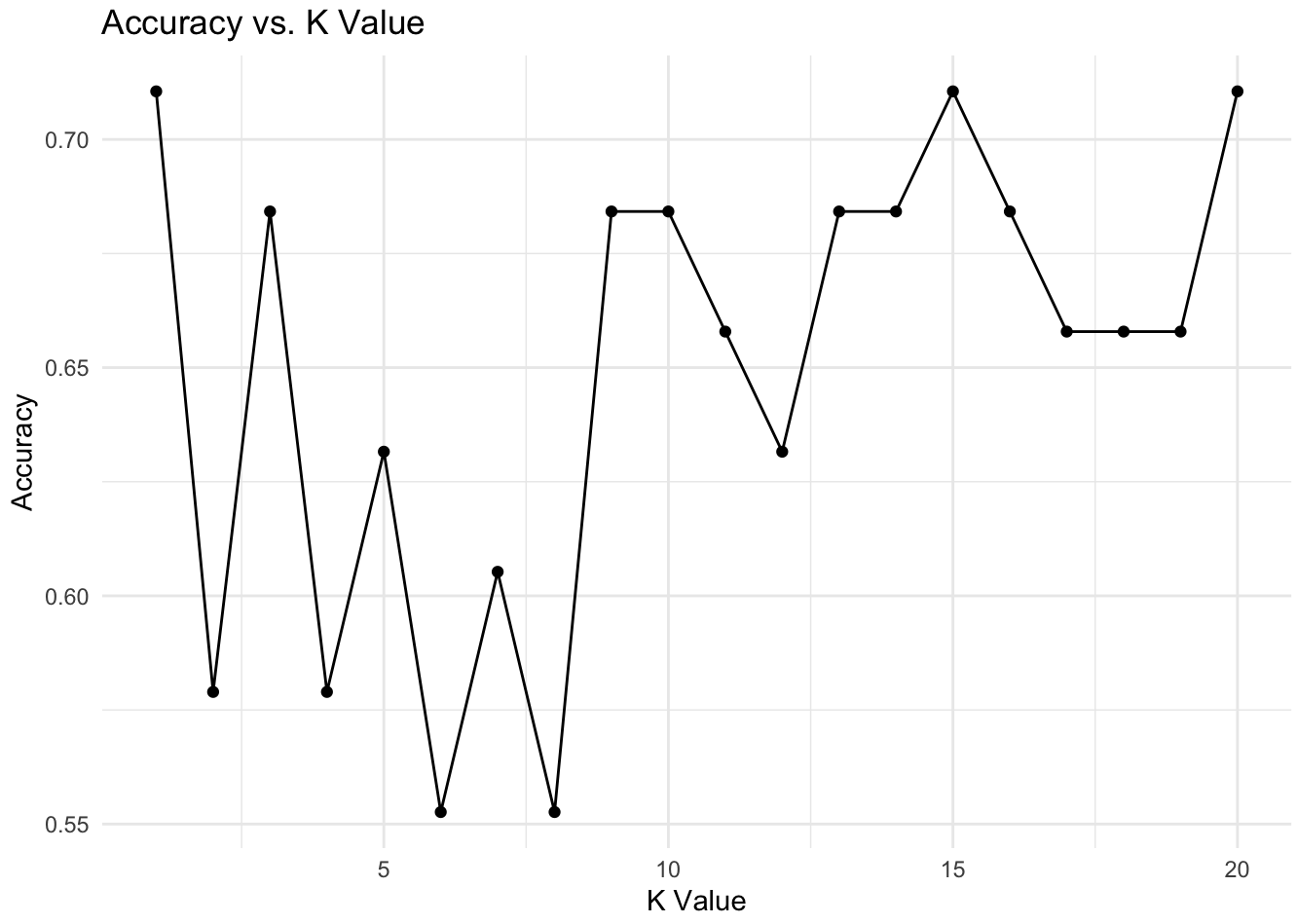
```
library(class)

set.seed(12)

# Initialize a vector to store accuracy for each k
accuracy_vector <- numeric(20)

# Loop over k values from 1 to 20
for (k in 1:20) {
  # Use knn to predict species on the validation set
  predicted_income <- knn(train_data[, -4], validation_data[, -4], train_data$other_info
rmation, k = k)

  # Calculate accuracy for this k
  accuracy <- sum(predicted_income == validation_data$other_information) / length(valida
tion_data$other_information)

  # Store the accuracy in the accuracy_vector
  accuracy_vector[k] <- accuracy
  cat("Accuracy for k =", k, ":", accuracy, "\n")
}
```

```
## Accuracy for k = 1 : 0.7105263
## Accuracy for k = 2 : 0.5789474
## Accuracy for k = 3 : 0.6842105
## Accuracy for k = 4 : 0.5789474
## Accuracy for k = 5 : 0.6315789
## Accuracy for k = 6 : 0.5526316
## Accuracy for k = 7 : 0.6052632
## Accuracy for k = 8 : 0.5526316
## Accuracy for k = 9 : 0.6842105
## Accuracy for k = 10 : 0.6842105
## Accuracy for k = 11 : 0.6578947
## Accuracy for k = 12 : 0.6315789
## Accuracy for k = 13 : 0.6842105
## Accuracy for k = 14 : 0.6842105
## Accuracy for k = 15 : 0.7105263
## Accuracy for k = 16 : 0.6842105
## Accuracy for k = 17 : 0.6578947
## Accuracy for k = 18 : 0.6578947
## Accuracy for k = 19 : 0.6578947
## Accuracy for k = 20 : 0.7105263
```

```
# Find the highest accuracy and its corresponding k
best_accuracy <- max(accuracy_vector)
best_k <- which(accuracy_vector == best_accuracy)

cat("The highest accuracy is", best_accuracy, "and it occurs for k =", best_k, "\n")
```

```
## The highest accuracy is 0.7105263 and it occurs for k = 1 15 20
```

```
# graph accuracy and k value
df <- data.frame(k = 1:20, accuracy = accuracy_vector)

ggplot(df, aes(x = k, y = accuracy)) +
  geom_line() +
  geom_point() +
  labs(title = "Accuracy vs. K Value", x = "K Value", y = "Accuracy") +
  theme_minimal()
```




```
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
set.seed(12)

#using k=15
predicted_demographic <- knn(train_data[, -4], test_data[, -4], train_data$other_informa
tion, k = 15)

print(table(predicted_demographic,test_data$other_information))
```

```
##
## predicted_demographic Minority-owned N/A
##        Minority-owned             17   4
##        N/A                         7  10
```

```
print(mean(predicted_demographic==test_data$other_information))
```

```
## [1] 0.7105263
```

# Random forest model

```
# Set seed for reproducibility
set.seed(12)

# use same cleaned dataframe from knn
women_income_subset_rf <- women_income_knn[,c("distance","angle","estimate_families_medi
an_income_dollars","other_information","professional_services", "entertainment_culture",
"beauty_wellness", "creative_economy", "retail", "services", "food", "healthcare", "educ
ation")]

# Split data into 60% training and 40% temporary from the total number of rows
train_set_indices_rf <- sample(1:nrow(women_income_subset_rf), 0.6 * nrow(women_income_s
ubset_rf), replace = FALSE)
train_data_rf <- women_income_subset_rf[train_set_indices_rf, ]
temp_data_rf <- women_income_subset_rf[-train_set_indices_rf, ]

# Split temp_data by 50% to get 20% valid and test data each
test_set_indices_rf <- sample(1:nrow(temp_data_rf), 0.5 * nrow(temp_data_rf), replace =
FALSE)
test_data_rf <- temp_data_rf[test_set_indices_rf, ]
validation_data_rf <- temp_data_rf[-test_set_indices_rf, ]


train_data_rf$other_information <- as.factor(train_data_rf$other_information)
test_data_rf$other_information <- as.factor(test_data_rf$other_information)
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:psych':
##
##     outlier
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
# set seed
set.seed(12)

#Random Forest model
rf <- randomForest(train_data_rf$other_information ~ distance + angle + estimate_familie
s_median_income_dollars + professional_services + entertainment_culture + beauty_wellnes
s + creative_economy + retail + services + food + healthcare + education,
                   data=train_data_rf,
                   mtry=12,
                   importance=TRUE)
rf_pred <- predict(rf, test_data_rf)
accuracy <- sum(rf_pred == test_data_rf$other_information) / nrow(test_data_rf)
print(accuracy)
```

```
## [1] 0.6052632
```

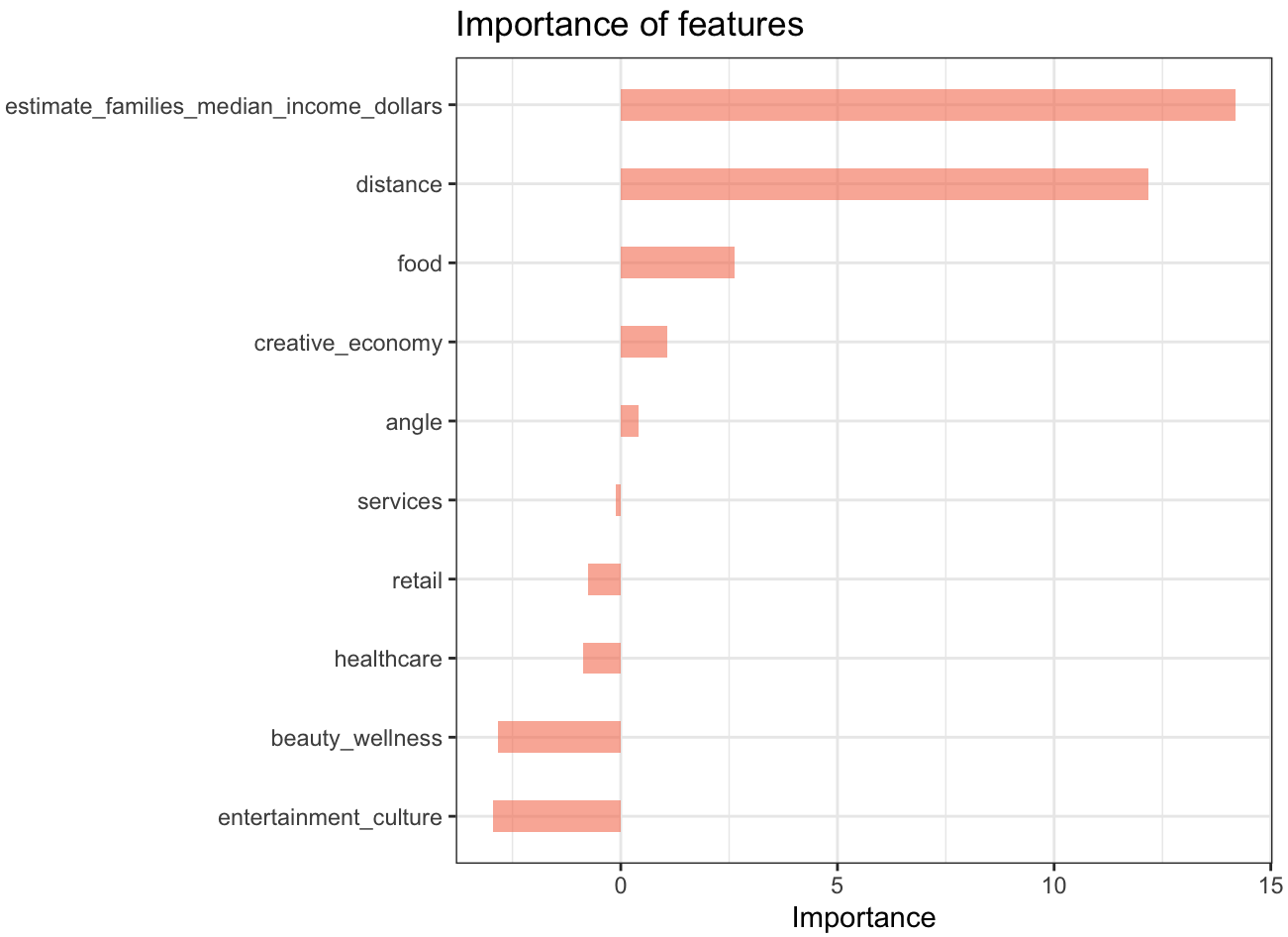
```
# look at most important variables
importance_vars <- importance(rf)
top_vars <- rownames(importance_vars[order(-importance_vars[,1]),])[1:10]
print(top_vars)
```

```
##  [1] "estimate_families_median_income_dollars"
##  [2] "distance"
##  [3] "food"
##  [4] "creative_economy"
##  [5] "angle"
##  [6] "services"
##  [7] "retail"
##  [8] "healthcare"
##  [9] "beauty_wellness"
## [10] "entertainment_culture"
```

```
# Assuming rf is a random forest model object from which you can extract feature importa
nce
importance_vars <- importance(rf)
top_vars <- rownames(importance_vars[order(-importance_vars[,1]),])[1:10]
print(top_vars)
```

```
##  [1] "estimate_families_median_income_dollars"
##  [2] "distance"
##  [3] "food"
##  [4] "creative_economy"
##  [5] "angle"
##  [6] "services"
##  [7] "retail"
##  [8] "healthcare"
##  [9] "beauty_wellness"
## [10] "entertainment_culture"
```

```
# Load necessary libraries
library(forcats)
library(ggplot2)

# You should replace `sample(1:10, 10)` with actual importance scores
data <- data.frame(
  name = top_vars,
  val = importance_vars[order(-importance_vars[,1]),1][1:10]
)

# Reorder and plot the data
ggplot(data, aes(x=fct_reorder(name, val), y=val)) +
  geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
  coord_flip() +
  labs(title = "Importance of features", x="", y="Importance") +
  theme_bw()
```

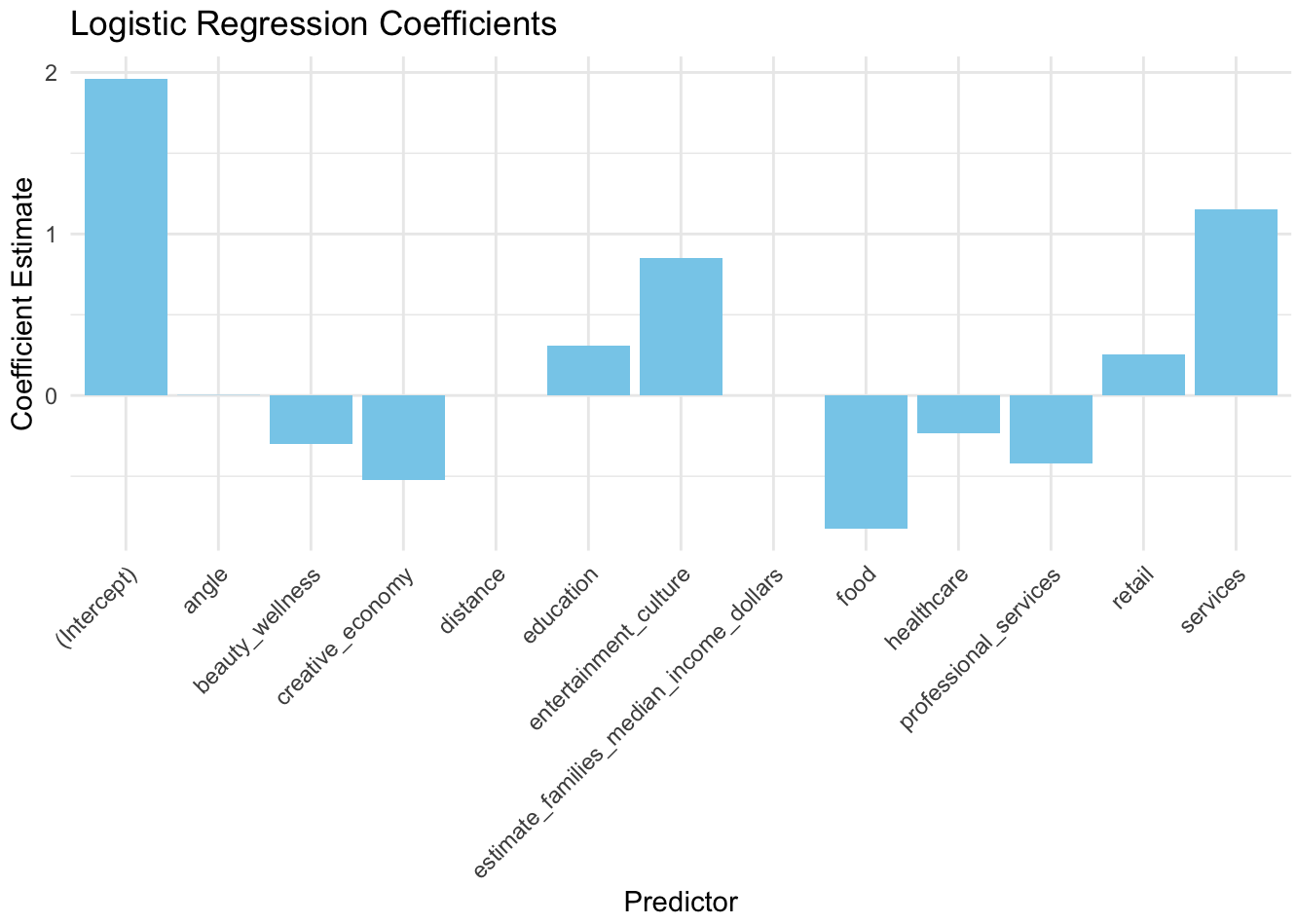



# Logistic regression model

```
set.seed(12)
# use women_income_knn dataframe because formatted for model
# fit the model
glm.fits <- glm(minority ~ distance + angle + estimate_families_median_income_dollars +
professional_services + entertainment_culture + beauty_wellness + creative_economy + ret
ail + services + food + healthcare + education,
                data = women_income_knn,
                family = binomial)
summary(glm.fits)
```

```
##
## Call:
## glm(formula = minority ~ distance + angle + estimate_families_median_income_dollars +
##     professional_services + entertainment_culture + beauty_wellness +
##     creative_economy + retail + services + food + healthcare +
##     education, family = binomial, data = women_income_knn)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             1.961e+00  5.361e-01   3.658 0.000255
## distance                               -8.499e-06  1.925e-05  -0.442 0.658785
## angle                                   1.990e-03  3.057e-03   0.651 0.514992
## estimate_families_median_income_dollars -1.073e-05  2.507e-06  -4.281 1.86e-05
## professional_services                  -4.233e-01  4.145e-01  -1.021 0.307181
## entertainment_culture                   8.535e-01  9.830e-01   0.868 0.385241
## beauty_wellness                        -2.978e-01  6.691e-01  -0.445 0.656289
## creative_economy                       -5.246e-01  5.149e-01  -1.019 0.308268
## retail                                  2.534e-01  4.853e-01   0.522 0.601560
## services                                1.151e+00  1.142e+00   1.008 0.313493
## food                                   -8.224e-01  5.352e-01  -1.537 0.124398
## healthcare                             -2.356e-01  5.349e-01  -0.440 0.659610
## education                               3.073e-01  4.753e-01   0.646 0.517993
##
## (Intercept)                            ***
## distance
## angle
## estimate_families_median_income_dollars ***
## professional_services
## entertainment_culture
## beauty_wellness
## creative_economy
## retail
## services
## food
## healthcare
## education
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 258.49  on 187  degrees of freedom
## Residual deviance: 224.34  on 175  degrees of freedom
## AIC: 250.34
##
## Number of Fisher Scoring iterations: 4
```

```r
# Load necessary libraries
library(ggplot2)

# Get model coefficients and convert to data frame
coefficients <- coef(summary(glm.fits))
coef_df <- as.data.frame(coefficients)

# Reset row names to create a variable column
coef_df$Variable <- rownames(coef_df)
rownames(coef_df) <- NULL

# Plot using ggplot2
ggplot(coef_df, aes(x = Variable, y = Estimate)) +
  geom_bar(stat = "identity", position = "dodge", fill = "skyblue") +
  theme_minimal() +
  labs(x = "Predictor", y = "Coefficient Estimate", title = "Logistic Regression Coeffic
ients") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels by 45
degrees
```



```
# prdictions and accuracy
predictions <- predict(glm.fits, type = "response")

predicted_class <- ifelse(predictions > 0.5, 1, 0)


confusion_matrix <- table(predicted_class, women_income_knn$minority)
print(confusion_matrix)
```

```
##
## predicted_class  0  1
##               0 46 25
##               1 38 79
```

```
accuracy <- sum(predicted_class == women_income_knn$minority) / nrow(women_income_knn)
print(accuracy)
```

```
## [1] 0.6648936
```

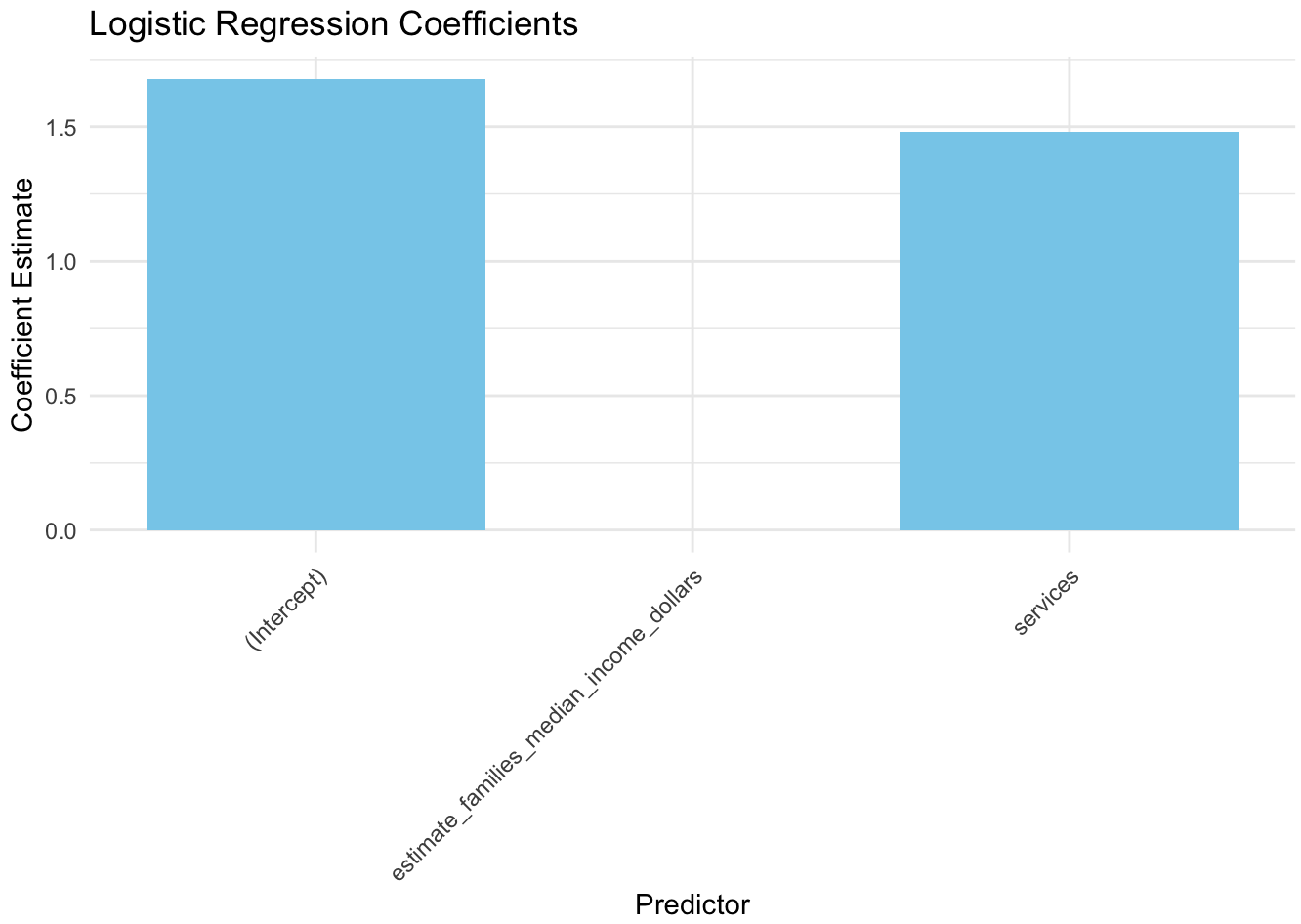
```
set.seed(12)

glm.start <- glm(minority ~ 1, data = women_income_knn, family = binomial)

#forward selection, using AIC
glm.forward <- step(glm.start, scope = list(lower = glm.start, upper = glm.fits),
                    direction = "forward")
```

```
## Start:  AIC=260.49
## minority ~ 1
##
##                                          Df Deviance    AIC
## + estimate_families_median_income_dollars  1   233.67 237.67
## + services                                 1   254.47 258.47
## + food                                     1   254.97 258.97
## + education                                1   255.84 259.83
## + professional_services                    1   256.46 260.46
## <none>                                         258.49 260.49
## + distance                                 1   257.56 261.56
## + angle                                    1   257.57 261.57
## + entertainment_culture                    1   257.70 261.70
## + creative_economy                         1   257.90 261.90
## + beauty_wellness                          1   258.27 262.27
## + retail                                   1   258.34 262.34
## + healthcare                               1   258.49 262.49
##
## Step:  AIC=237.67
## minority ~ estimate_families_median_income_dollars
##
##                          Df Deviance    AIC
## + services                1   231.25 237.25
## <none>                        233.67 237.67
## + food                    1   231.92 237.92
## + education               1   232.35 238.35
## + angle                   1   232.66 238.66
## + creative_economy        1   232.74 238.74
## + entertainment_culture   1   232.76 238.76
## + professional_services   1   232.82 238.82
## + retail                  1   233.04 239.04
## + distance                1   233.53 239.53
## + beauty_wellness         1   233.61 239.61
## + healthcare              1   233.66 239.66
##
## Step:  AIC=237.25
## minority ~ estimate_families_median_income_dollars + services
##
##                          Df Deviance    AIC
## <none>                        231.25 237.25
## + food                    1   229.76 237.76
## + education               1   229.76 237.76
## + angle                   1   230.41 238.41
## + retail                  1   230.41 238.41
## + entertainment_culture   1   230.51 238.51
## + creative_economy        1   230.53 238.53
## + professional_services   1   230.70 238.70
## + distance                1   231.04 239.04
## + beauty_wellness         1   231.23 239.23
## + healthcare              1   231.25 239.25
```

```
summary(glm.forward)
```

```
##
## Call:
## glm(formula = minority ~ estimate_families_median_income_dollars +
##     services, family = binomial, data = women_income_knn)
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           1.677e+00  3.698e-01   4.535 5.75e-06
## estimate_families_median_income_dollars -1.057e-05  2.313e-06  -4.571 4.86e-06
## services                              1.482e+00  1.103e+00   1.343    0.179
##
## (Intercept)                           ***
## estimate_families_median_income_dollars ***
## services
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 258.49  on 187  degrees of freedom
## Residual deviance: 231.25  on 185  degrees of freedom
## AIC: 237.25
##
## Number of Fisher Scoring iterations: 4
```

```
# Get model coefficients and convert to data frame
coefficients <- coef(summary(glm.forward))
coef_df <- as.data.frame(coefficients)

# Reset row names to create a variable column
coef_df$Variable <- rownames(coef_df)
rownames(coef_df) <- NULL

# Plot using ggplot2
ggplot(coef_df, aes(x = Variable, y = Estimate)) +
  geom_bar(stat = "identity", position = "dodge", fill = "skyblue") +
  theme_minimal() +
  labs(x = "Predictor", y = "Coefficient Estimate", title = "Logistic Regression Coeffic
ients") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels by 45
degrees
```




```
# prdictions and accuracy
predictions <- predict(glm.forward, type = "response")

predicted_class <- ifelse(predictions > 0.5, 1, 0)


confusion_matrix <- table(predicted_class, women_income_knn$minority)
print(confusion_matrix)
```

```
##
## predicted_class  0  1
##               0 44 25
##               1 40 79
```

```
accuracy <- sum(predicted_class == women_income_knn$minority) / nrow(women_income_knn)
print(accuracy)
```

```
## [1] 0.6542553
```