title: “Foundations Project” output: html_document date: “2023-10-25” —

Set up

```
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── 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 be
come errors
```

```
library(sf)
```

```
## Linking to GEOS 3.11.2, GDAL 3.7.2, PROJ 9.3.0; sf_use_s2() is TRUE
```

```
library(terra)
```

```
## terra 1.7.55
##
## 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)
```

```
## Warning: package 'tmap' was built under R version 4.3.2
```

```
## Breaking News: tmap 3.x is retiring. Please test v4, e.g. with
## remotes::install_github('r-tmap/tmap')
```

```
library(flexmix)
```

```
## Warning: package 'flexmix' was built under R version 4.3.2
```

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

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

```
## Warning: package 'geosphere' was built under R version 4.3.2
```

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 duplic
ated 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" ~ "Minority
-owned",
                                       other_information == "Immigrant-owned, N/A" ~ "Immigra
nt-owned",
                                       .default = other_information)) # this is to standardiz
e and combine forms
```

```
## Rows: 265 Columns: 10
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (9): Business Name, Business Type, Physical Location/Address, Street add...
## dbl (1): Business Zipcode
##
## 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 faster
access to ACS tables in the future, run this function with `cache_table = TRUE`. You only nee
d 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%
  |
  |=                                                                     |   2%
  |
  |==                                                                    |   2%
  |
  |==                                                                    |   3%
  |
  |===                                                                   |   4%
  |
  |===                                                                   |   5%
  |
  |====                                                                  |   5%
  |
  |====                                                                  |   6%
  |
  |=====                                                                 |   7%
  |
  |=====                                                                 |   8%
  |
  |======                                                                |   8%
  |
  |======                                                                |   9%
  |
  |=======                                                               |   9%
  |
  |=======                                                               |  10%
  |
  |========                                                              |  11%
  |
  |========                                                              |  12%
  |
  |=========                                                             |  12%
  |
  |=========                                                             |  13%
  |
  |==========                                                            |  14%
  |
  |==========                                                            |  15%
  |
  |===========                                                           |  15%
  |
  |===========                                                           |  16%
  |
  |============                                                          |  17%
  |
  |=============                                                         |  18%
```

```
  |
  |=============                                                         |  19%
  |
  |==============                                                        |  20%
  |
  |==============                                                        |  21%
  |
  |===============                                                       |  21%
  |
  |===============                                                       |  22%
  |
  |================                                                      |  22%
  |
  |================                                                      |  23%
  |
  |=================                                                     |  24%
  |
  |=================                                                     |  25%
  |
  |==================                                                    |  25%
  |
  |==================                                                    |  26%
  |
  |===================                                                   |  27%
  |
  |===================                                                   |  28%
  |
  |====================                                                  |  28%
  |
  |====================                                                  |  29%
  |
  |=====================                                                 |  30%
  |
  |======================                                                |  31%
  |
  |======================                                                |  32%
  |
  |=======================                                               |  33%
  |
  |========================                                              |  34%
  |
  |========================                                              |  35%
  |
  |=========================                                             |  36%
  |
  |==========================                                            |  37%
  |
  |==========================                                            |  38%
  |
  |===========================                                           |  39%
  |
```

```
|============================                                          |  40%
|
|============================                                          |  41%
|
|=============================                                         |  42%
|
|==============================                                        |  43%
|
|==============================                                        |  44%
|
|===============================                                       |  44%
|
|===============================                                       |  45%
|
|================================                                      |  46%
|
|=================================                                     |  47%
|
|==================================                                    |  48%
|
|==================================                                    |  49%
|
|===================================                                   |  49%
|
|===================================                                   |  50%
|
|====================================                                  |  51%
|
|====================================                                  |  52%
|
|=====================================                                 |  52%
|
|=====================================                                 |  53%
|
|======================================                                |  54%
|
|======================================                                |  55%
|
|=======================================                               |  55%
|
|=======================================                               |  56%
|
|========================================                              |  57%
|
|========================================                              |  58%
|
|=========================================                             |  58%
|
|=========================================                             |  59%
|
|==========================================                            |  59%
```

```
|
|==========================================                            |  60%
|
|===========================================                           |  61%
|
|===========================================                           |  62%
|
|============================================                          |  62%
|
|============================================                          |  63%
|
|=============================================                         |  64%
|
|=============================================                         |  65%
|
|==============================================                        |  65%
|
|==============================================                        |  66%
|
|===============================================                       |  67%
|
|===============================================                       |  68%
|
|================================================                      |  68%
|
|================================================                      |  69%
|
|=================================================                     |  70%
|
|=================================================                     |  71%
|
|==================================================                    |  71%
|
|==================================================                    |  72%
|
|===================================================                   |  73%
|
|====================================================                  |  74%
|
|====================================================                  |  75%
|
|=====================================================                 |  75%
|
|=====================================================                 |  76%
|
|======================================================                |  77%
|
|======================================================                |  78%
|
|=======================================================               |  78%
|
```

```
  |=======================================================               |  79%
  |
  |========================================================              |  79%
  |
  |========================================================              |  80%
  |
  |=========================================================             |  81%
  |
  |=========================================================             |  82%
  |
  |==========================================================            |  82%
  |
  |==========================================================            |  83%
  |
  |===========================================================           |  84%
  |
  |===========================================================           |  85%
  |
  |============================================================          |  85%
  |
  |============================================================          |  86%
  |
  |=============================================================         |  87%
  |
  |==============================================================        |  88%
  |
  |==============================================================        |  89%
  |
  |===============================================================       |  89%
  |
  |===============================================================       |  90%
  |
  |================================================================      |  91%
  |
  |================================================================      |  92%
  |
  |=================================================================     |  92%
  |
  |=================================================================     |  93%
  |
  |==================================================================    |  94%
  |
  |===================================================================   |  95%
  |
  |===================================================================   |  96%
  |
  |====================================================================  |  97%
  |
  |===================================================================== |  98%
  |
  |===================================================================== |  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 200 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 Market"
women$street_address[women$street_address == "Commercial Wharf"] <- "47 Commercial Wharf"
```

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"] <- -71.05516

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
```

Now we have 197 values without geocode. After checking them all over, many are virtual businesses and therefore do not have store fronts. Many businesses also appear to be out of date as their website does not show up.

There was only one business (F2 Fitness Wellness), that we could not find an address for but appeared to be up and running and in person. Their website lists DC as the address, but the street address they gave does not line up with the zip code they provided. So we exclude 197 businesses total, but 196 due to lack of physical presence.

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 = (12:13), 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" | business_t
ype == "Real Estate Broker/Owner" | business_type == "Website Design" |business_type == "Prof
essional Services, Coach women entrepreneurs start a business"|business_type == "Creative Eco
nomy, Professional Services, Advertising, Marketing, Branding"|business_type == "Development
and Construction"|business_type == "Financial Services, Healthcare, Professional Services"|bu
siness_type == "Financial Services, Healthcare, Professional Services, Consulting" |business_
type == "Construction"|business_type == "Financial Services" |business_type == "Creative Econ
omy, Professional Services, Retail, Creative Agency"|business_type == "Professional Services,
Residential and Commercial Real Estate Sales and Leasing"|business_type == "Retail, Interior
Design & Construction Project Management"|business_type == "Communications and Public Affair
s" |business_type == "Education, Financial Services"|business_type == "Professional Services,
Executive Search/Recruiting/Human Resources Consulting"|business_type == "Clean-tech/Green-te
ch, Education, Healthcare, Professional Services, Cleaning industry"|business_type == "Creati
ve Economy, Professional Services"|business_type == "Real Estate"|business_type == "Life Coac
hing"|business_type == "Architecture"|business_type == "Coaching"|business_type ==  "Professi
onal Services, software testing and data analysis - we can work with any business, any indust
ry"|business_type == "Technology"|business_type == "Professional Services, Real Estate Broker
age"|business_type == "Clean-tech/Green-tech, Professional Services"|business_type == "Profes
sional Services, Women's Empowerment Groups virtual & in person"|business_type == "Profession
al Services, Real Estate"|business_type == "Clean-tech/Green-tech, Professional Services, ELE
CTRICAL AND FIRE ALARM SERVICES"|business_type == "Professional Services, Business Launch & L
ife Alignment Coaching"|business_type == "Clean-tech/Green-tech, Manufacturing, 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, Manufa
cturing, Professional Services, Restaurant & Catering, Retail, Technology, Tourism"|business_
type == "Clean-tech/Green-tech"|business_type == "Education, Financial Services, Professional
Services"|business_type == "Education, Professional Services, Self Development, Communication
Skills Coaching"|business_type == "Education, Financial Services, Professional Services, COAC
HING AND MENTORING"|business_type == "Creative Economy, Financial Services, Professional Serv
ices"|business_type == "Legal and Investigative Group"|business_type == "Professional Service
s, NOTARY PUBLIC, counseling multi services"|business_type == "Healthcare, Professional Servi
ces, Social Service", 1, 0),
         entertainment_culture = ifelse(business_type ==  "Tourism, Food, Culture, History of
Boston's Chinatown"|business_type ==  "Recreational sports and social events"|business_type =
=  "Professional Services, Entertainment"|business_type ==  "Tourism" |business_type ==  "Bio
-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Education, Financial Service
s, Food and Beverage, Healthcare, Manufacturing, Professional Services, Restaurant & Caterin
g, Retail, Technology, Tourism"|business_type ==  "E-commerce Natural Skin Care Brand" |busin
ess_type ==  "Provide event staff"|business_type ==  "Kid entertainment"|business_type ==  "E
vents"|business_type ==  "Party Rental Company and Event Space", 1, 0),
         beauty_wellness = ifelse(business_type ==  "apperal ,beauty and health supplies"|bus
iness_type ==  "Salon"|business_type ==  "Professional Services, Hair and Makeup"|business_ty
pe ==  "Pet Care"|business_type ==  "Makeup Artistry"|business_type ==  "Beauty Salon"|busine
ss_type ==  "Professional Services, Hair Salon"|business_type ==  "Wellness"|business_type ==
"Health and wellness"|business_type ==  "Healthcare, Holistic Wellness"|business_type ==  "He
althcare, Hair care"|business_type ==   "Education, Professional Services, Retail, hair salo
n"|business_type ==  "Creative Economy, Professional Services, Beauty Services"|business_type
==  "Esthetician"|business_type ==  "Fashion/Beauty"|business_type ==  "Spa"|business_type ==
"I formulate plant based skin care" |business_type == "Hair Salon", 1, 0),
```

```
         creative_economy = ifelse(business_type ==  "Creative Economy"|business_type == "Cre
ative Economy, Food and Beverage, Restaurant & Catering"|business_type == "Creative Economy,
Food and Beverage, Restaurant & Catering"|business_type == "Creative Economy, Professional Se
rvices, Advertising, Marketing, Branding"|business_type == "Creative Economy, Professional Se
rvices, Retail, Creative Agency" |business_type ==  "Retail, Interior Design & Construction P
roject Management"|business_type == "Bio-tech & Life Sciences, Creative Economy, Healthcare"|
business_type == "Art"|business_type == "Creative Agency" |business_type == "Creative Econom
y, Professional Services"|business_type == "Broadcast Media"|business_type == "Creative Econo
my, Education, Retail"|business_type == "Creative Economy, Contemplative + Healing Arts"|busi
ness_type == "Creative Economy, Retail"|business_type == "Interior Design Services"|business_
type ==  "Food and Beverage, Retail, Florist"|business_type == "Creative Economy, Professiona
l Services, Beauty Services"|business_type == "Creative Economy, Graphic Design" |business_ty
pe == "Creative Economy, Manufacturing"|business_type == "Creative Economy, Professional Serv
ices, Photography"|business_type == "Creative Economy, Education, Professional Services"|busi
ness_type == "Creative Economy, Market Research, Strategy and Design"|business_type ==  "Crea
tive Economy, Professional Services, Retail", 1, 0),
         retail = ifelse(business_type ==  "apperal ,beauty and health supplies"|business_typ
e == "Retail, Handmade"|business_type == "Retail" |business_type == "Food and Beverage, Retai
l"|business_type == "Restaurant & Catering, Retail"|business_type == "Creative Economy, Profe
ssional Services, Retail, Creative Agency"|business_type == "Retail, Interior Design & Constr
uction Project Management"|business_type == "Creative Economy, Education, Retail"|business_ty
pe ==  "Education, Retail" |business_type == "Creative Economy, Retail" |business_type == "Bi
o-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Education, Financial Service
s, Food and Beverage, Healthcare, Manufacturing, Professional Services, Restaurant & Caterin
g, Retail, Technology, Tourism"|business_type == "Education, Professional Services, Retail, h
air salon"|business_type == "Education, Food and Beverage"|business_type == "Creative Econom
y, Professional Services, Retail", 1, 0),
         services = ifelse(business_type ==  "Cleaning"|business_type == "Clean-tech/Green-te
ch, Cleaning Company"|business_type == "Building Services-Cleaning"|business_type == "Towin
g"|business_type ==  "Green Cleaning (Residential"|business_type == "Clean-tech/Green-tech, E
ducation, Healthcare, Personal maid" |business_type == "Clean-tech/Green-tech, Education, Hea
lthcare, Professional Services, Cleaning industry"|business_type == "Janitorial"|business_typ
e == "Pet Care"|business_type ==  "Clothing alteration and dry cleaning"|business_type == "Co
nstruction Painting"|business_type == "Provide event staff", 1, 0),
         food = ifelse(business_type ==  "Food and Beverage"|business_type == "Restaurant & C
atering, Retail"|business_type == "Restaurant & Catering"|business_type == "Food and Beverag
e, Restaurant & Catering"|business_type == "Food and Beverage, Retail"|business_type == "Crea
tive Economy, Food and Beverage, Restaurant & Catering"|business_type == "Tourism, Food, Cult
ure, History of Boston's Chinatown"|business_type == "Creative Economy, Food and Beverage, Re
staurant & Catering"|business_type == "Extra Virgin Olive Oil"|business_type == "Bio-tech & L
ife Sciences, Clean-tech/Green-tech, Creative Economy, Education, Financial Services, Food an
d Beverage, Healthcare, Manufacturing, Professional Services, Restaurant & Catering, Retail,
Technology, Tourism"|business_type == "CPG Food Company - Frozen Meal Bites for Kids"|busines
s_type == "Food and Beverage, Retail, Florist"|business_type == "Dog Bakery", 1, 0),
         healthcare = ifelse(business_type ==  "Healthcare"|business_type == "Health & Wellne
ss"|business_type == "Medical spa/ Day Spa"|business_type == "Forensic Science"|business_type
== "Education, Healthcare, Retail, Fitness/Wellness"|business_type == "Education, Healthcare,
Professional Services, Heath & Wellness"|business_type == "Suicide Prevention - Military and
for Spanish speakers" |business_type == "Education, Healthcare, Professional Services"|busine
ss_type == "Financial Services, Healthcare, Professional Services"|business_type == "Financia
l Services, Healthcare, Professional Services, Consulting"|business_type ==  "Bio-tech & Life
```

```
Sciences, Creative Economy, Healthcare"|business_type == "Clean-tech/Green-tech, Education, H
ealthcare, Personal maid"|business_type == "Clean-tech/Green-tech, Education, Healthcare, Pro
fessional Services, Cleaning industry" |business_type == "Bio-tech & Life Sciences, Clean-tec
h/Green-tech, Creative Economy, Education, Financial Services, Food and Beverage, Healthcare,
Manufacturing, Professional Services, Restaurant & Catering, Retail, Technology, Tourism"|bus
iness_type == "Health and wellness"|business_type == "Healthcare, Holistic Wellness" |busines
s_type ==  "Health and Fitness" |business_type ==  "Healthcare, Professional Services, Social
Service", 1, 0),
         education = ifelse(business_type ==  "Education, Professional Services, Technology,
Data science / predictive modeling"|business_type ==  "Education, Professional Services"|busi
ness_type ==  "Education, Nonprofit"|business_type ==  "Education"|business_type ==  "Educati
on, Professional Services, Non Profit"|business_type ==  "Education, Swim School"|business_ty
pe ==  "Education, Healthcare, Professional Services, Heath & Wellness"|business_type ==  "Ed
ucation, Healthcare, Retail, Fitness/Wellness"|business_type ==  "Education, Healthcare, Prof
essional Services" |business_type ==  "Education, Financial Services" |business_type ==  "Cle
an-tech/Green-tech, Education, Healthcare, Personal maid"|business_type ==  "Clean-tech/Green
-tech, Education, Healthcare, Professional Services, Cleaning industry"|business_type ==  "Cr
eative Economy, Education, Retail"|business_type ==  "Education, Retail"|business_type ==  "B
io-tech & Life Sciences, Clean-tech/Green-tech, Creative Economy, Education, Financial Servic
es, Food and Beverage, Healthcare, Manufacturing, Professional Services, Restaurant & Caterin
g, Retail, Technology, Tourism"|business_type ==  "Education, Financial Services, Professiona
l Services"|business_type ==  "Education, Professional Services, Self Development, Communicat
ion Skills Coaching" |business_type ==  "Education, Financial Services, Professional Service
s, COACHING AND MENTORING"|business_type ==  "Education, Professional Services, Retail, hair
salon"|business_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, Veteran-own
ed"] <- "Minority-owned"
```

```
# Join latitude and longitude columns go dataframe
women_income <- left_join(women_income, women_geocoded[c("cxy_lat", "cxy_lon", "business_nam
e")])
```

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

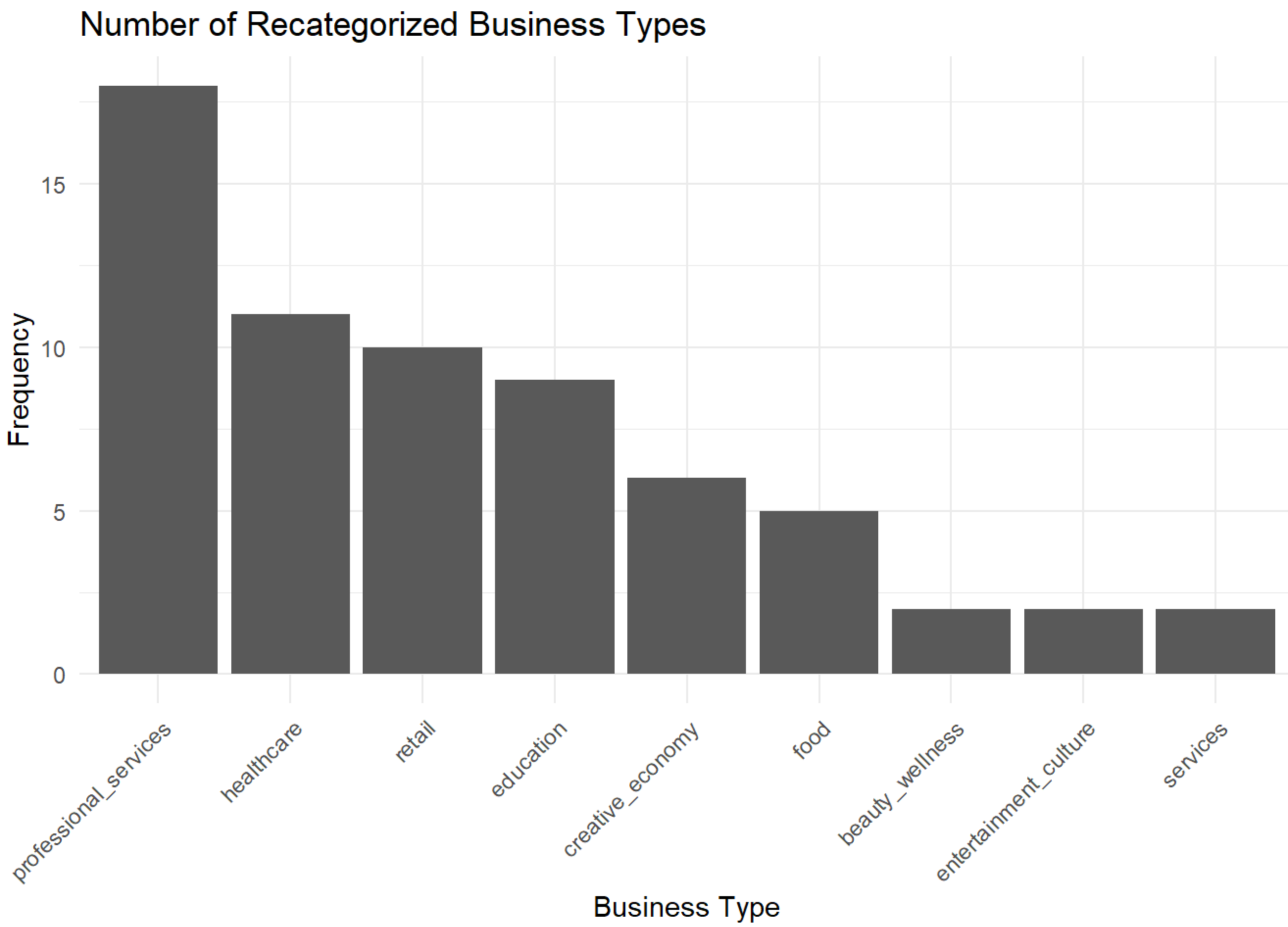
Calculate distance from center of Boston and

angle

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

Descriptive statistics

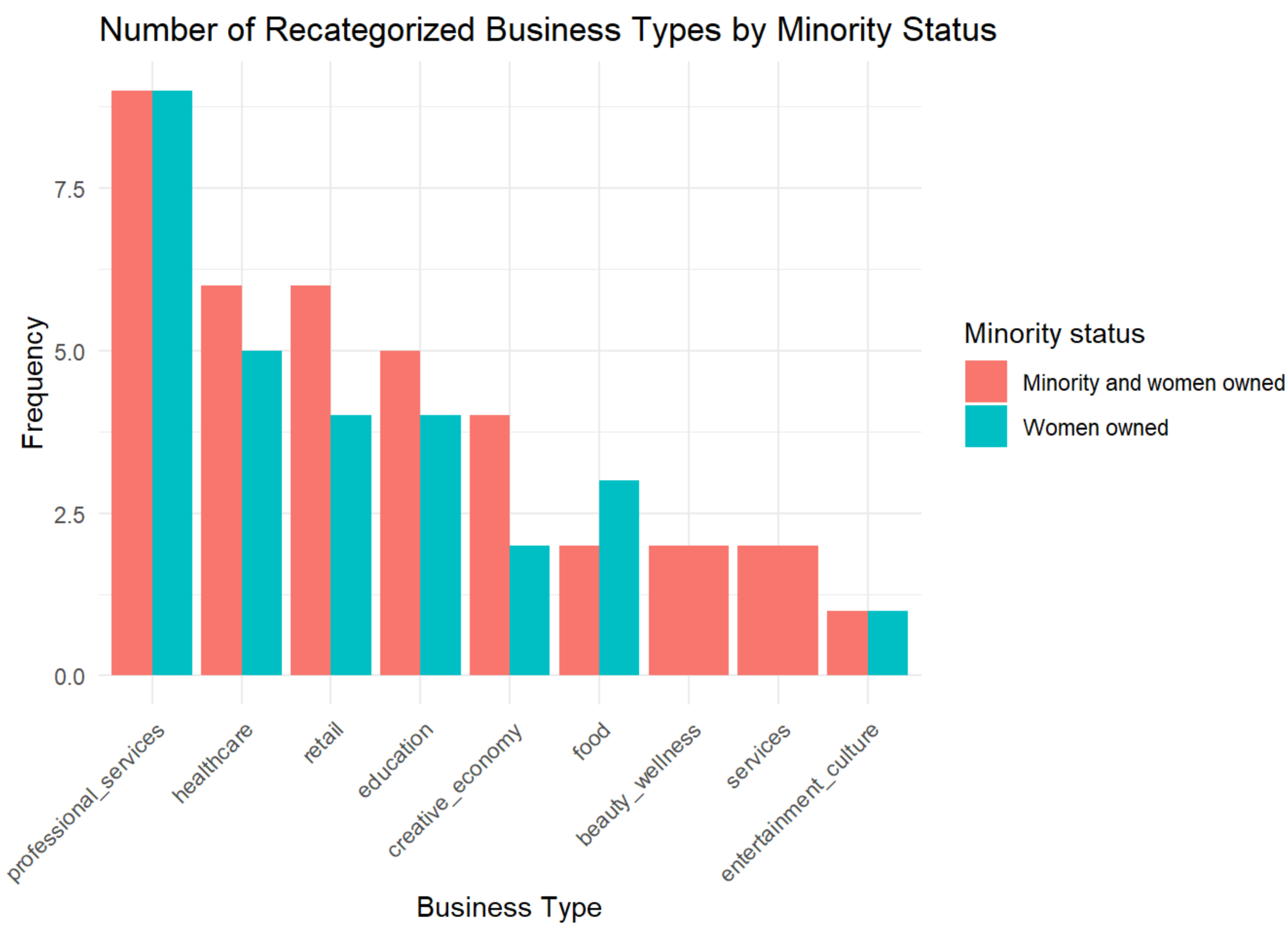
```
# Identify the most common recategorized business types
top_category_types <- women_income %>%
  st_drop_geometry() %>%
  count(professional_services, education, food, services, retail, creative_economy, entertain
ment_culture, beauty_wellness, healthcare) %>%
  pivot_longer(cols = "professional_services":"healthcare") %>%
  # 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(""))
```



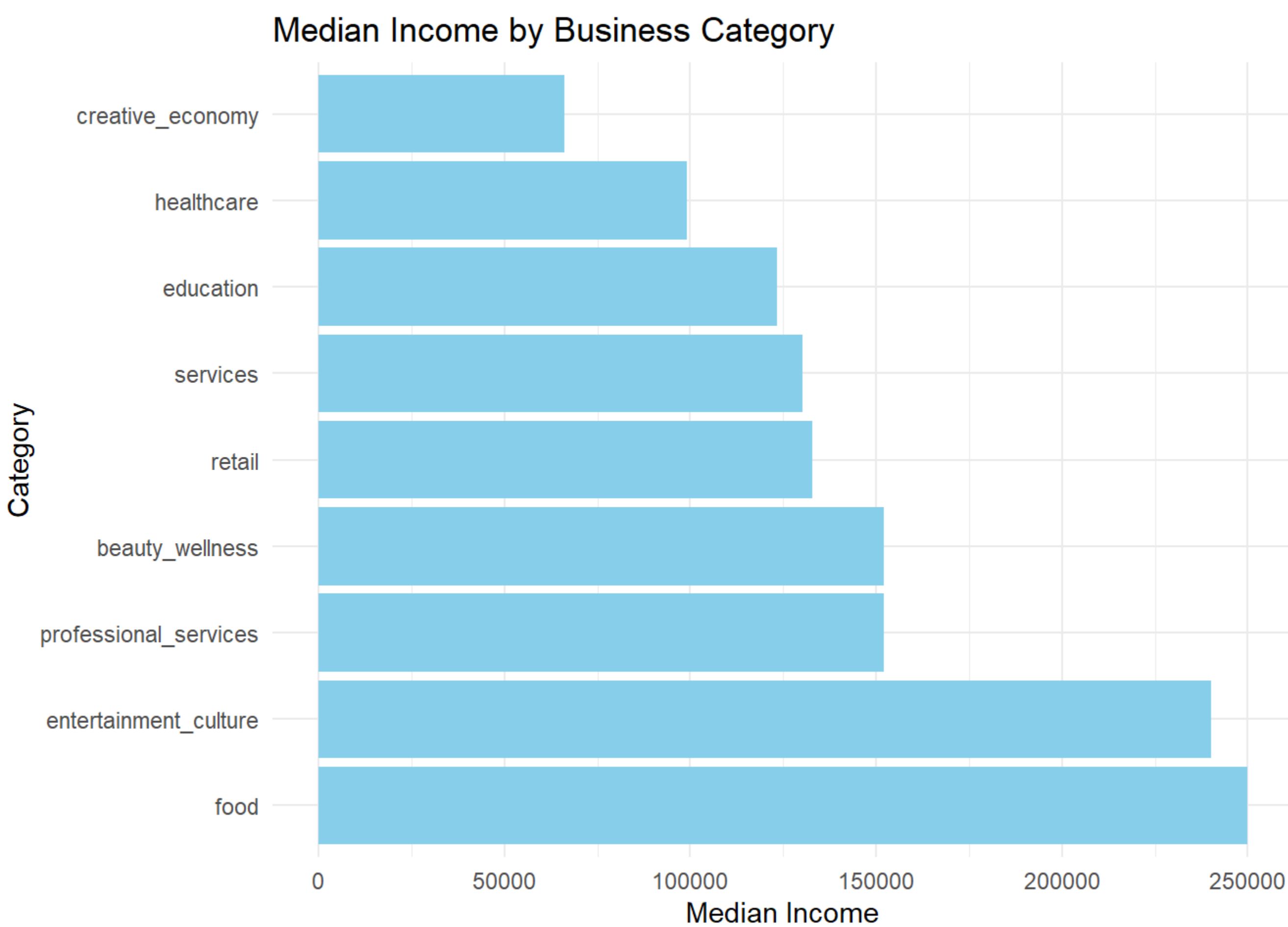

```
# Identify the most common recategorized business types
top_category_types_minority <- women_income %>%
  st_drop_geometry() %>%
  count(professional_services, education, food, services, retail, creative_economy, entertain
ment_culture, beauty_wellness, healthcare, minority) %>%
  pivot_longer(cols = "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 owne
d"))
```

```
## `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 status
`)) +
  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, entertai
nment_culture, beauty_wellness, healthcare, estimate_families_median_income_dollars) %>%
  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 Income")
```

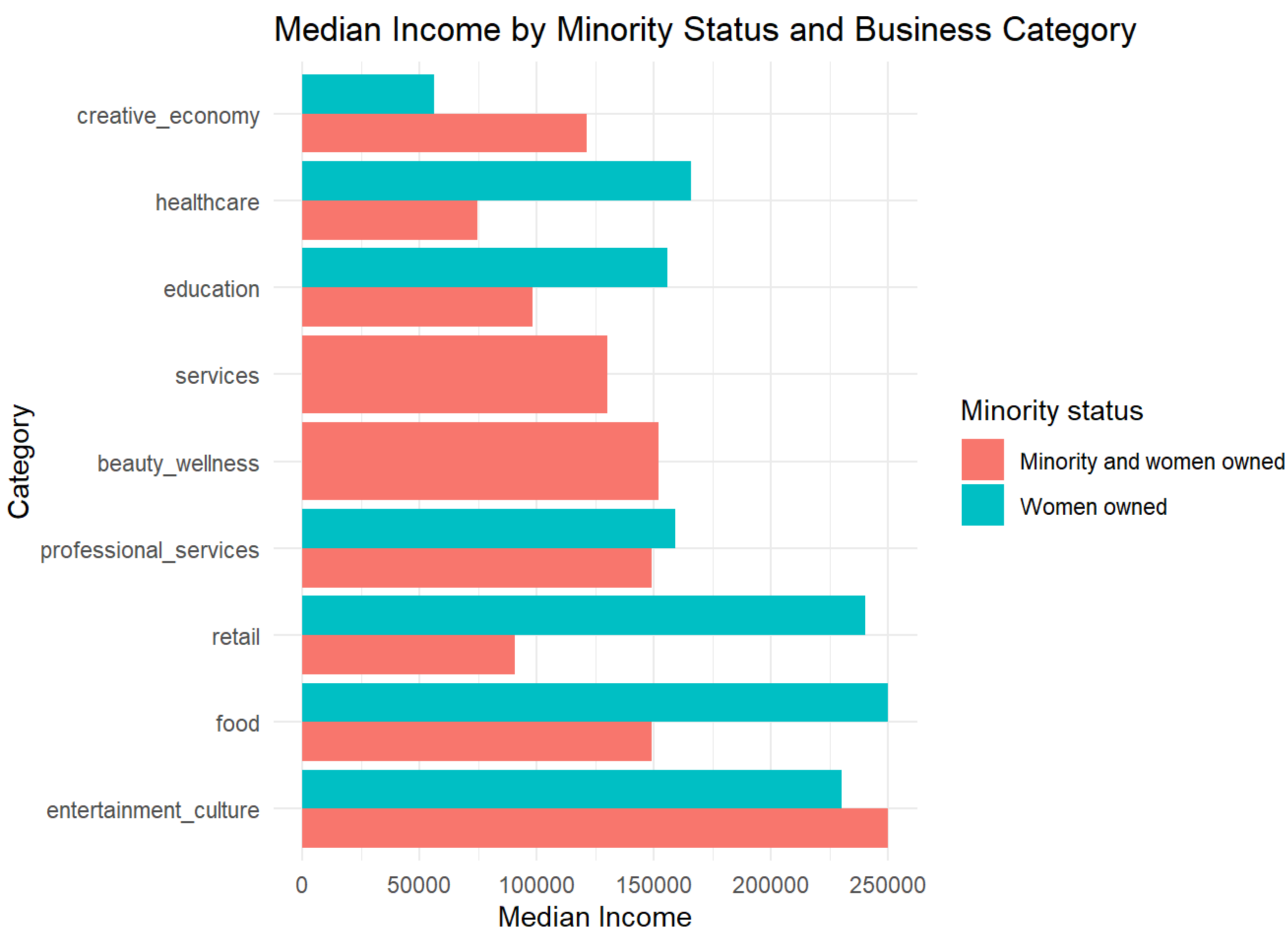


```
# Group data by minority/immigrant status and calculate median income
median_income_by_minority <- women_income %>%
  group_by(minority) %>%
  summarise(median_income = median(estimate_families_median_income_dollars, na.rm = TRUE)) %
>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority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

# Visualize the results
ggplot(median_income_by_minority, aes(x = as.factor(`Minority status`), y = median_income)) +
  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, entertai
nment_culture, beauty_wellness, healthcare, estimate_families_median_income_dollars, minorit
y) %>%
  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 = TRUE)) %
>%
  mutate(`Minority status` = ifelse(minority == 1, "Minority and women owned", "Women owne
d"))
```

```
## `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_income, fil
l = `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")
```

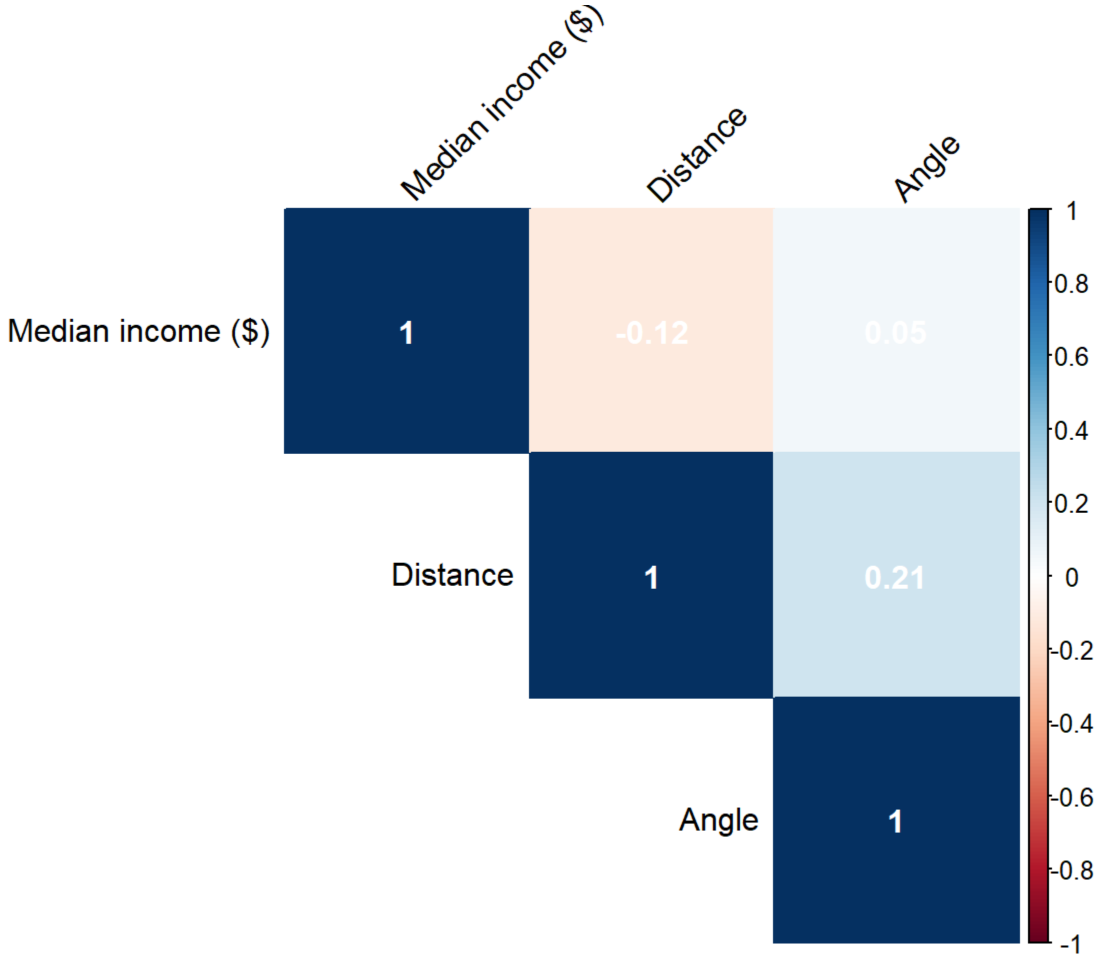



```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

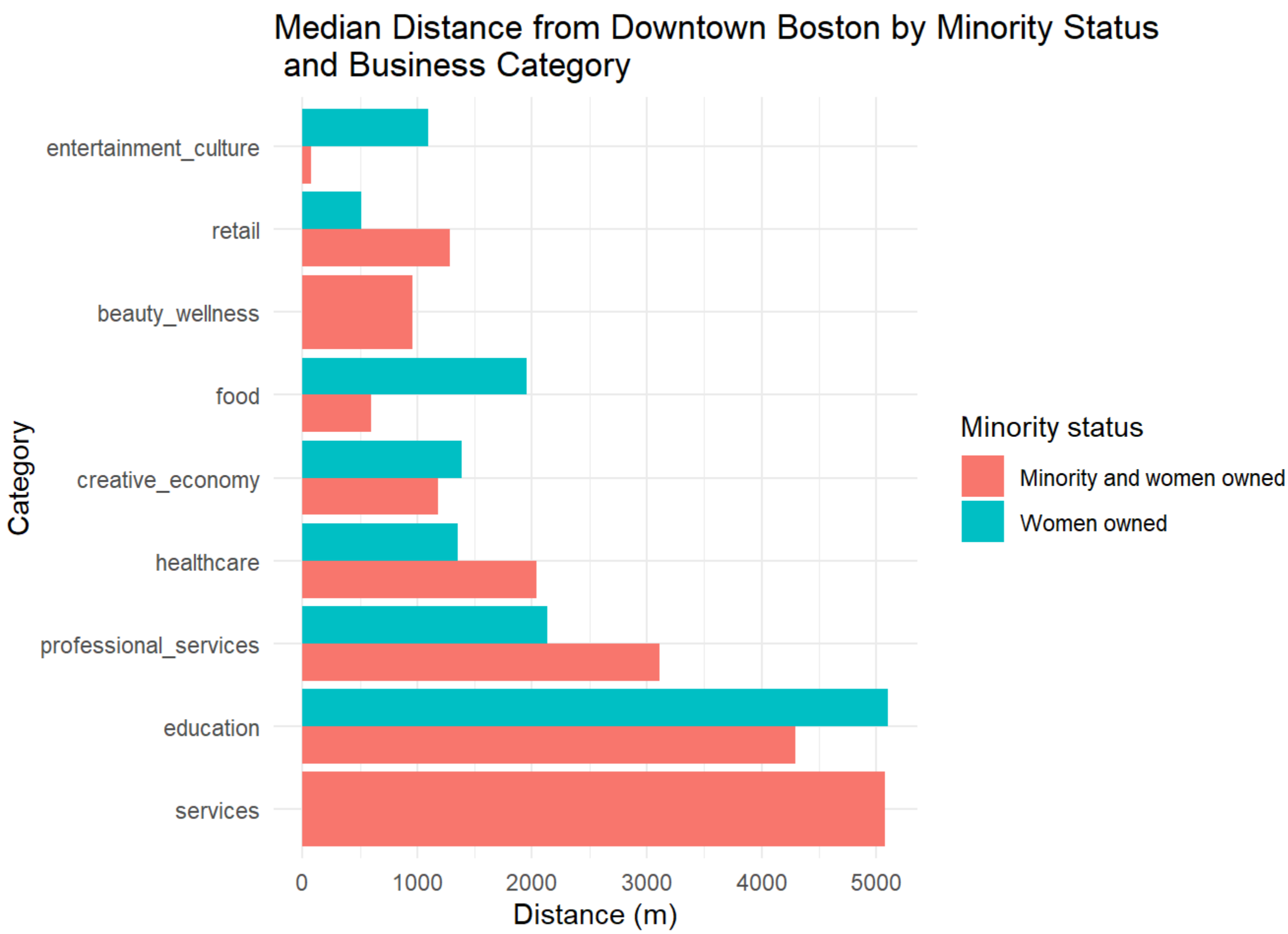
```
women_income_corrplot <- women_income %>%
  rename(`Median income ($)` = estimate_families_median_income_dollars,
         Distance = distance,
         Angle = angle)

# Correlation matrix for selected variables
correlation_matrix <- cor(women_income_corrplot[,c("Median income ($)", "Distance", "Angle")]
%>%
  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, entertai
nment_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 owne
d"))
```

```
## `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_distance,
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 Categ
ory", x = "Category", y = "Distance (m)")
```

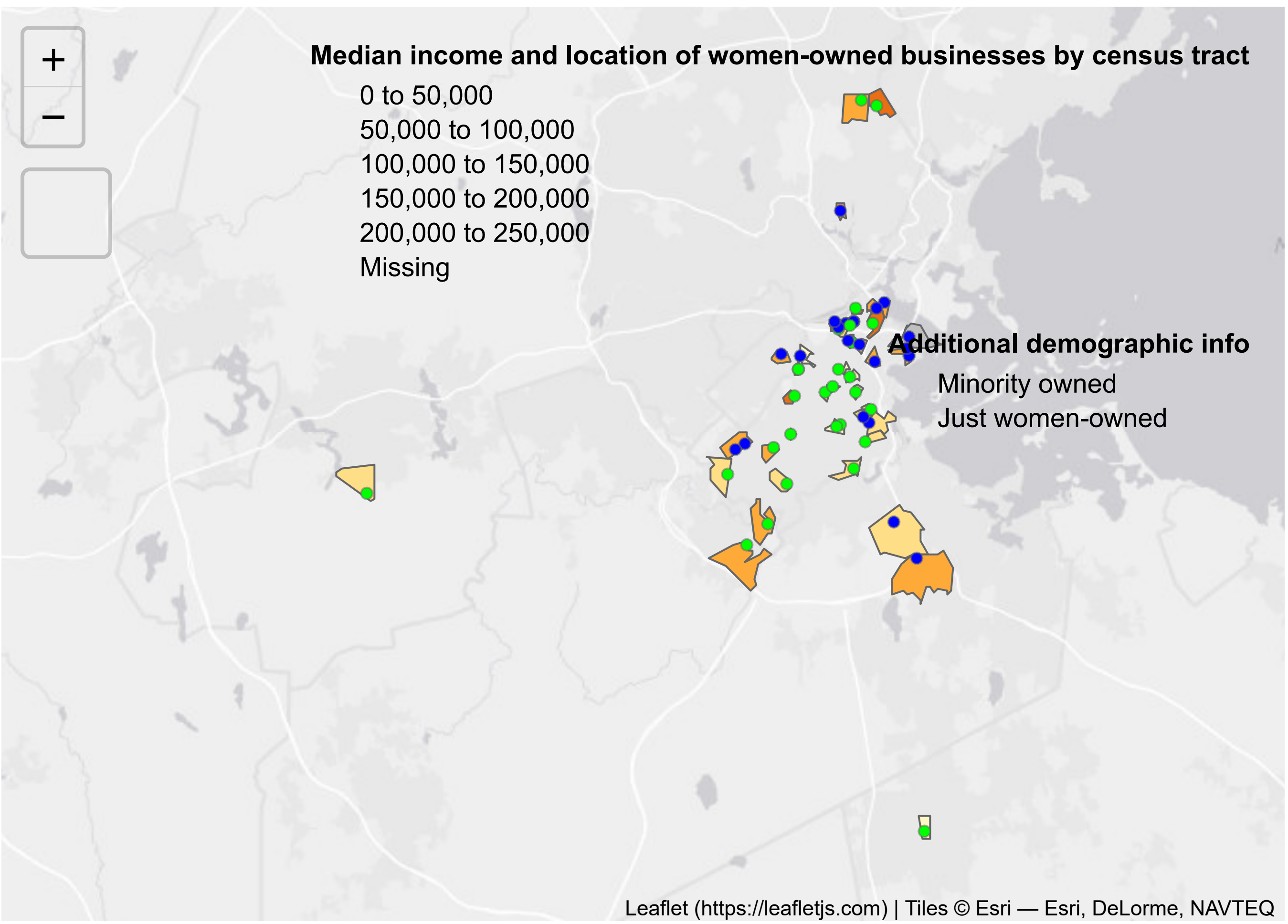


Map of income and location of businesses

```r
# 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 locati
on 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), labels =
c("Minority owned", "Just women-owned"), palette=c('green','blue'))
```

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

```r
map0
```



PCA of median income, distance, and angle

```r
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] 1.2257773 1.0445730 0.7296496
```

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

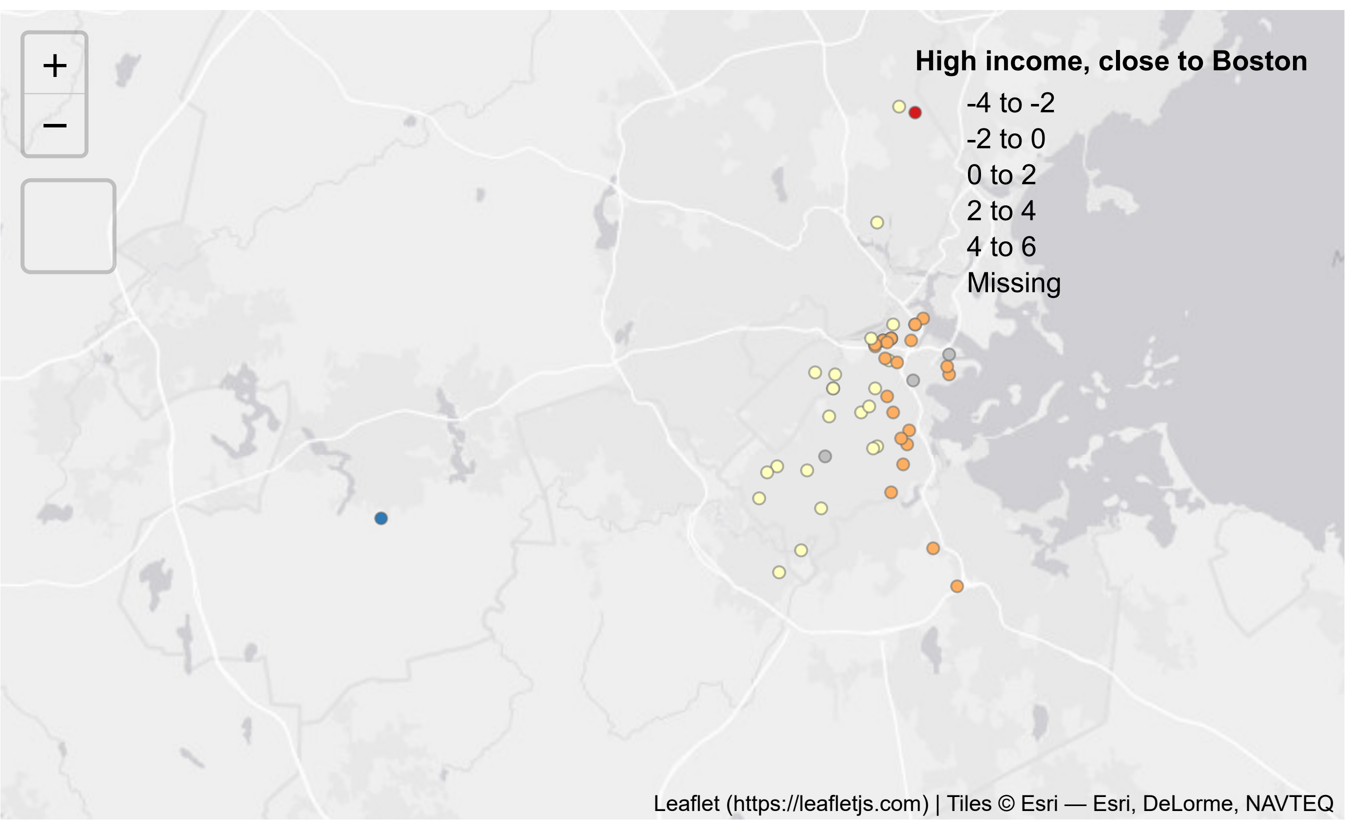
```
## estimate_families_median_income_dollars                                distance
##                                       1                                       1
##                                   angle
##                                       1
```

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

```
##
## Loadings:
##                                    PC1    PC2    PC3
## estimate_families_median_income_dollars -0.258  0.899  0.355
## distance                            0.810 -0.126  0.573
## angle                               0.709  0.470 -0.525
##
##                  PC1   PC2   PC3
## SS loadings    1.226 1.045 0.730
## Proportion Var 0.409 0.348 0.243
## Cumulative Var 0.409 0.757 1.000
```

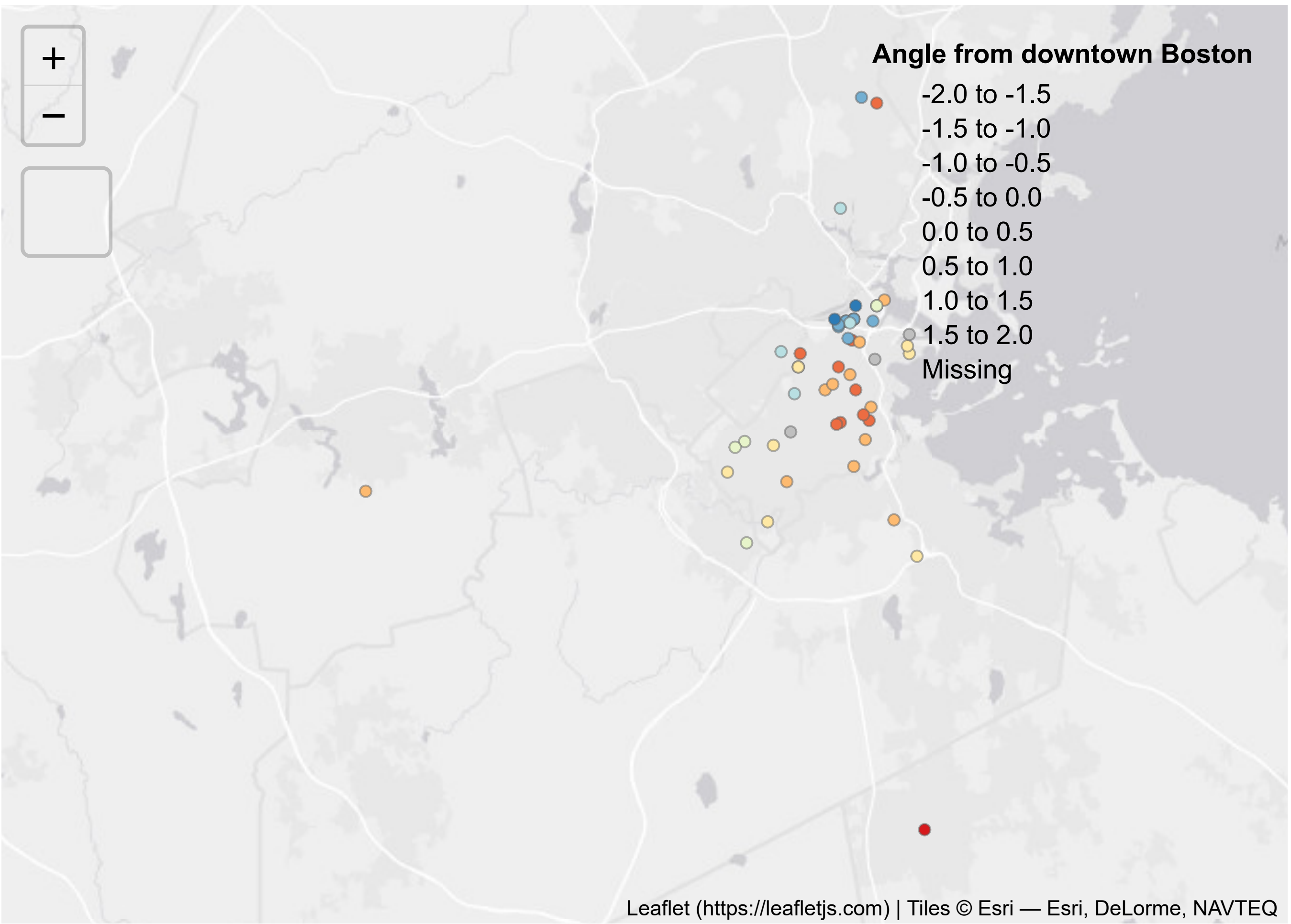
```
library(RColorBrewer)

#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 Bo
ston", palette = pc_palette)
map_pc1
```

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



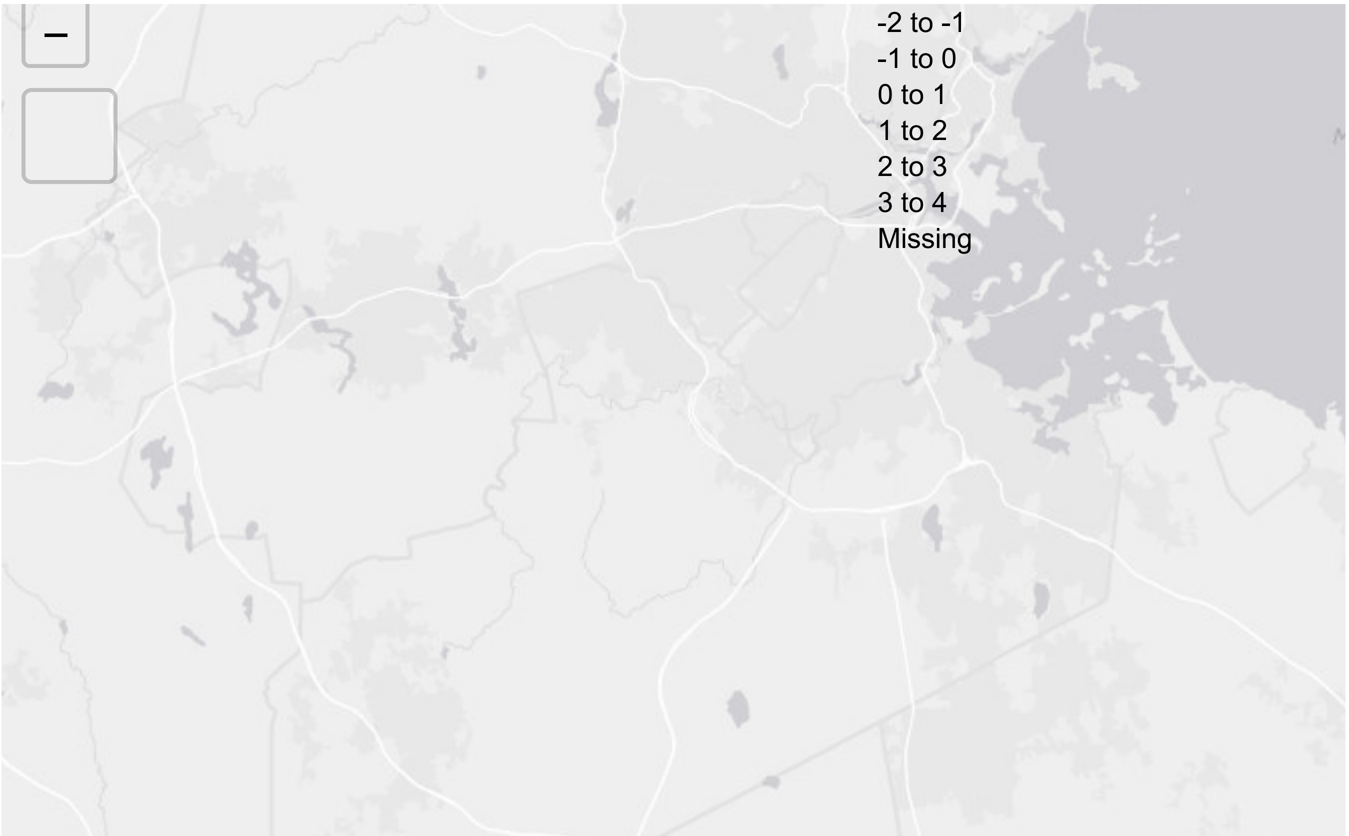
```
map_pc2 <- tm_shape(women_income_pca) +tm_dots(col="PC2", title="Angle from downtown Boston",
palette = pc_palette)
map_pc2
```

```
## Variable(s) "PC2" contains positive and negative values, so midpoint is set to 0. Set midp
oint = 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 midp
oint = 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-owned", "
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_dollar
s)
```

```
# 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_norm","othe
r_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_subset), rep
lace = 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: 52 × 4
## # Rowwise:
##    distance_norm[,1] angle_norm[,1] median_norm[,1] other_information
##                <dbl>          <dbl>           <dbl> <chr>
##  1           -0.452          -0.254         -1.26   Minority-owned
##  2           -0.536           2.76           0.108  Minority-owned
##  3           -0.0641         -0.630         -0.0387 N/A
##  4           -0.638          -0.427         -1.21   N/A
##  5           -0.329           0.545          1.10   Minority-owned
##  6            0.0368         -0.550         -1.54   Minority-owned
##  7           -0.226           0.606          1.36   N/A
##  8           -0.665          -0.445         -0.848  Minority-owned
##  9            0.642           0.180         -0.412  Minority-owned
## 10           -0.174          -1.11           0.0184 N/A
## # ℹ 42 more rows
```

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

```
## [1] 52  4
```

```
dim(train_data)
```

```
## [1] 31  4
```

```
dim(test_data)
```

```
## [1] 10  4
```

```
dim(validation_data)
```

```
## [1] 11  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
```

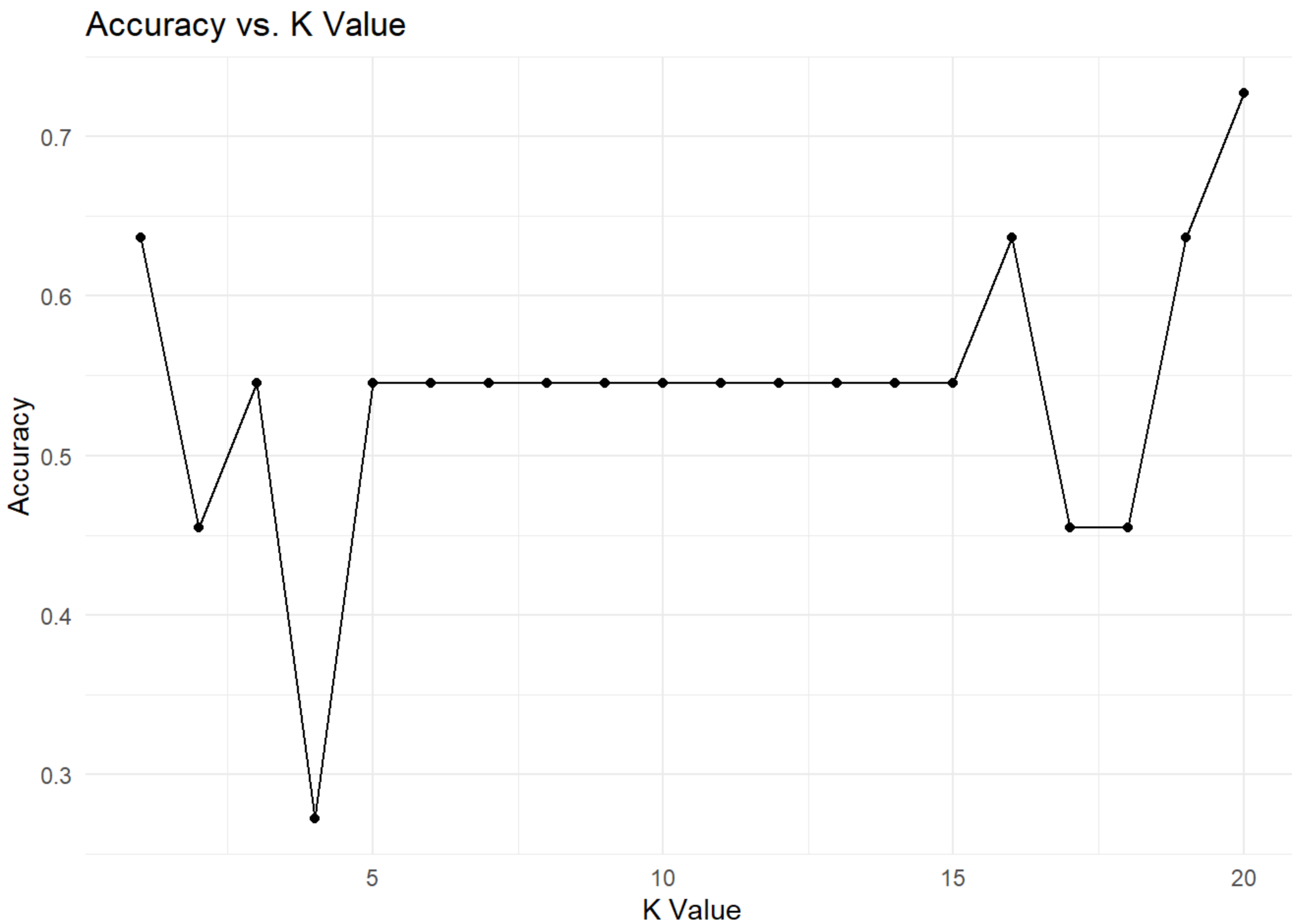
```
library(class)

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_informati
on, k = k)

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

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

```
## Accuracy for k = 1 : 0.6363636
## Accuracy for k = 2 : 0.4545455
## Accuracy for k = 3 : 0.5454545
## Accuracy for k = 4 : 0.2727273
## Accuracy for k = 5 : 0.5454545
## Accuracy for k = 6 : 0.5454545
## Accuracy for k = 7 : 0.5454545
## Accuracy for k = 8 : 0.5454545
## Accuracy for k = 9 : 0.5454545
## Accuracy for k = 10 : 0.5454545
## Accuracy for k = 11 : 0.5454545
## Accuracy for k = 12 : 0.5454545
## Accuracy for k = 13 : 0.5454545
## Accuracy for k = 14 : 0.5454545
## Accuracy for k = 15 : 0.5454545
## Accuracy for k = 16 : 0.6363636
## Accuracy for k = 17 : 0.4545455
## Accuracy for k = 18 : 0.4545455
## Accuracy for k = 19 : 0.6363636
## Accuracy for k = 20 : 0.7272727
```

```
# 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 0.7272727 and it occurs for k = 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)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
##
## 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_information,
k = 15)

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

```
##
## predicted_demographic Minority-owned N/A
##        Minority-owned              5   3
##        N/A                         1   1
```

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

```
## [1] 0.6
```

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_median_in
come_dollars","other_information","professional_services", "entertainment_culture", "beauty_w
ellness", "creative_economy", "retail", "services", "food", "healthcare", "education")]

# 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_subset
_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 = FALS
E)
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$other_information ~ distance + angle + estimate_families_med
ian_income_dollars + professional_services + entertainment_culture + beauty_wellness + creati
ve_economy + retail + services + food + healthcare + education,
                    data=train_data_rf,
                    mtry=12,
                    importance=TRUE)
rf_pred <- predict(rf, test_data_rf)
accuracy <- sum(rf_pred == test_data_rf$other_information) / nrow(test_data_rf)
print(accuracy)
```

```
## [1] 0.7
```

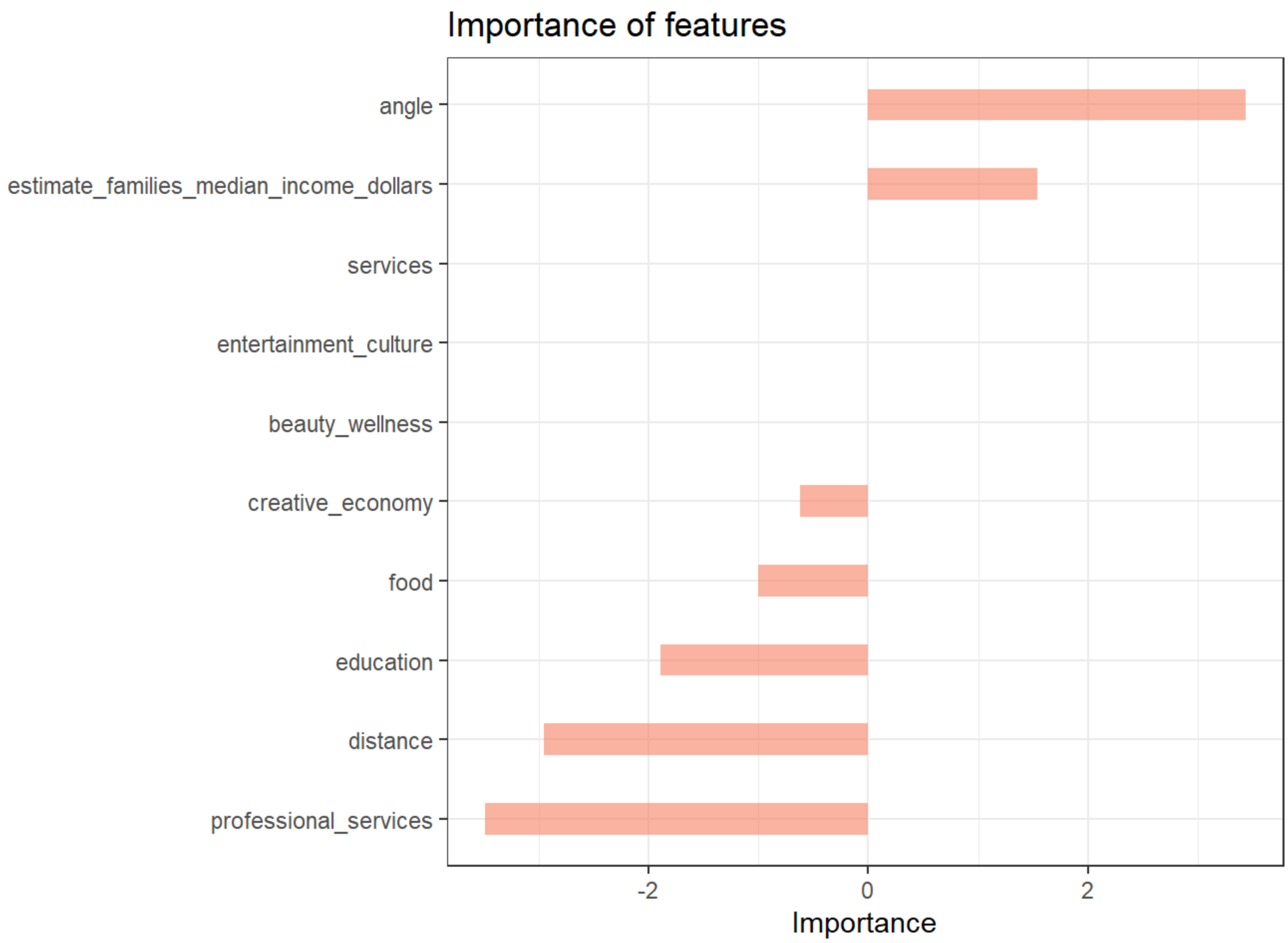
```
# Look at most important variables
importance_vars <- importance(rf)
top_vars <- rownames(importance_vars[order(-importance_vars[,1]),])[1:10]
print(top_vars)
```

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

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

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

```r
# 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()
```

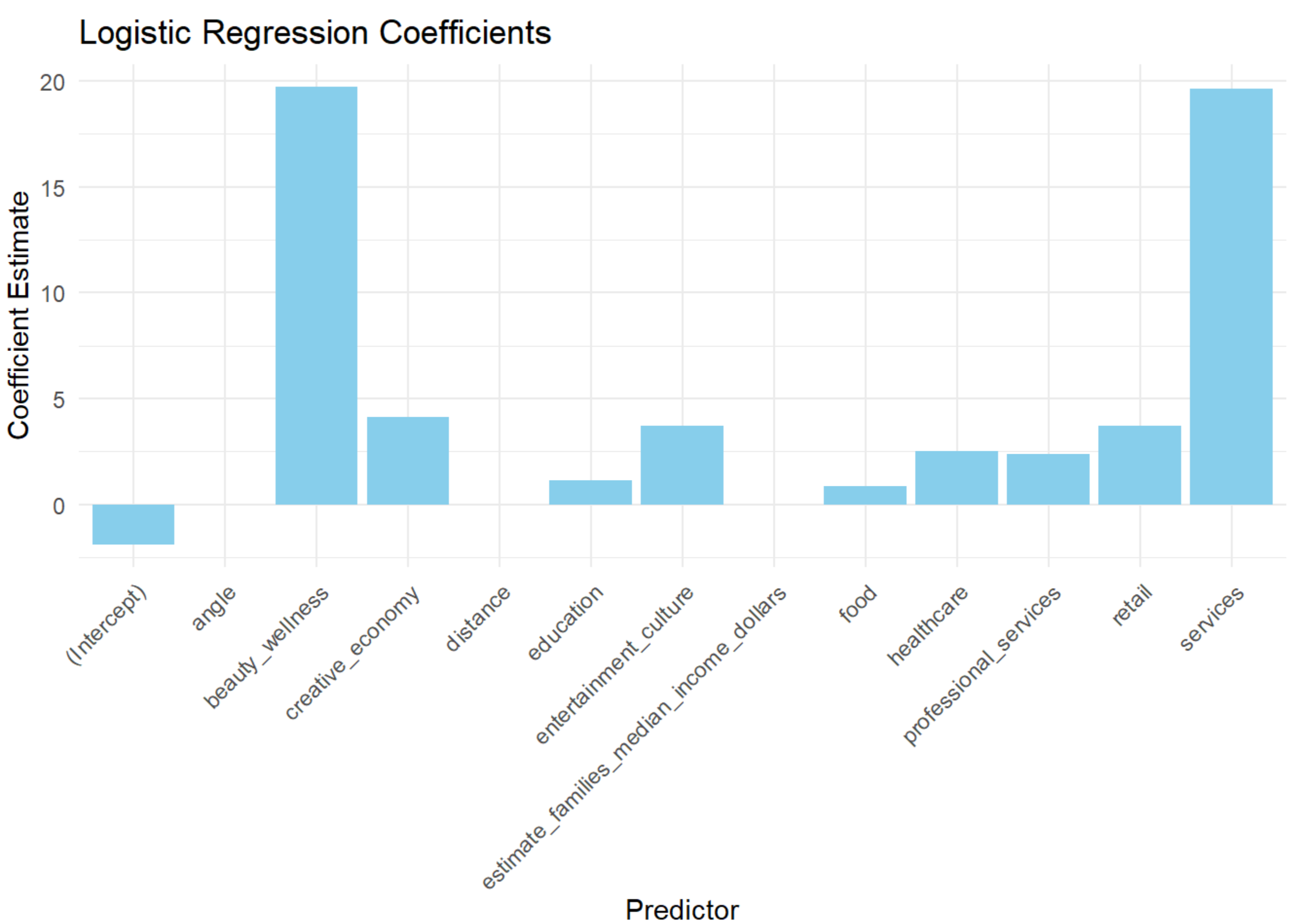


Logistic regression model

```
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 + profe
ssional_services + entertainment_culture + beauty_wellness + creative_economy + retail + serv
ices + 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)                            -1.883e+00  1.748e+00  -1.077   0.2814
## distance                                1.702e-04  1.802e-04   0.944   0.3449
## angle                                   1.246e-02  7.790e-03   1.599   0.1098
## estimate_families_median_income_dollars -1.093e-05  5.344e-06  -2.045   0.0409
## professional_services                   2.373e+00  1.509e+00   1.573   0.1158
## entertainment_culture                   3.696e+00  2.158e+00   1.713   0.0867
## beauty_wellness                         1.971e+01  2.081e+03   0.009   0.9924
## creative_economy                        4.119e+00  1.835e+00   2.244   0.0248
## retail                                  3.720e+00  1.815e+00   2.050   0.0404
## services                                1.963e+01  2.744e+03   0.007   0.9943
## food                                    8.479e-01  1.599e+00   0.530   0.5959
## healthcare                              2.508e+00  1.546e+00   1.623   0.1046
## education                               1.136e+00  1.460e+00   0.778   0.4365
## 
## (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: 71.393  on 51  degrees of freedom
## Residual deviance: 52.351  on 39  degrees of freedom
## AIC: 78.351
## 
## Number of Fisher Scoring iterations: 16
```

```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 Coefficient
s") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels by 45 degre
es
```



```
# 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 15  7
##               1  8 22
```

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

```
## [1] 0.7115385
```

```
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")
```

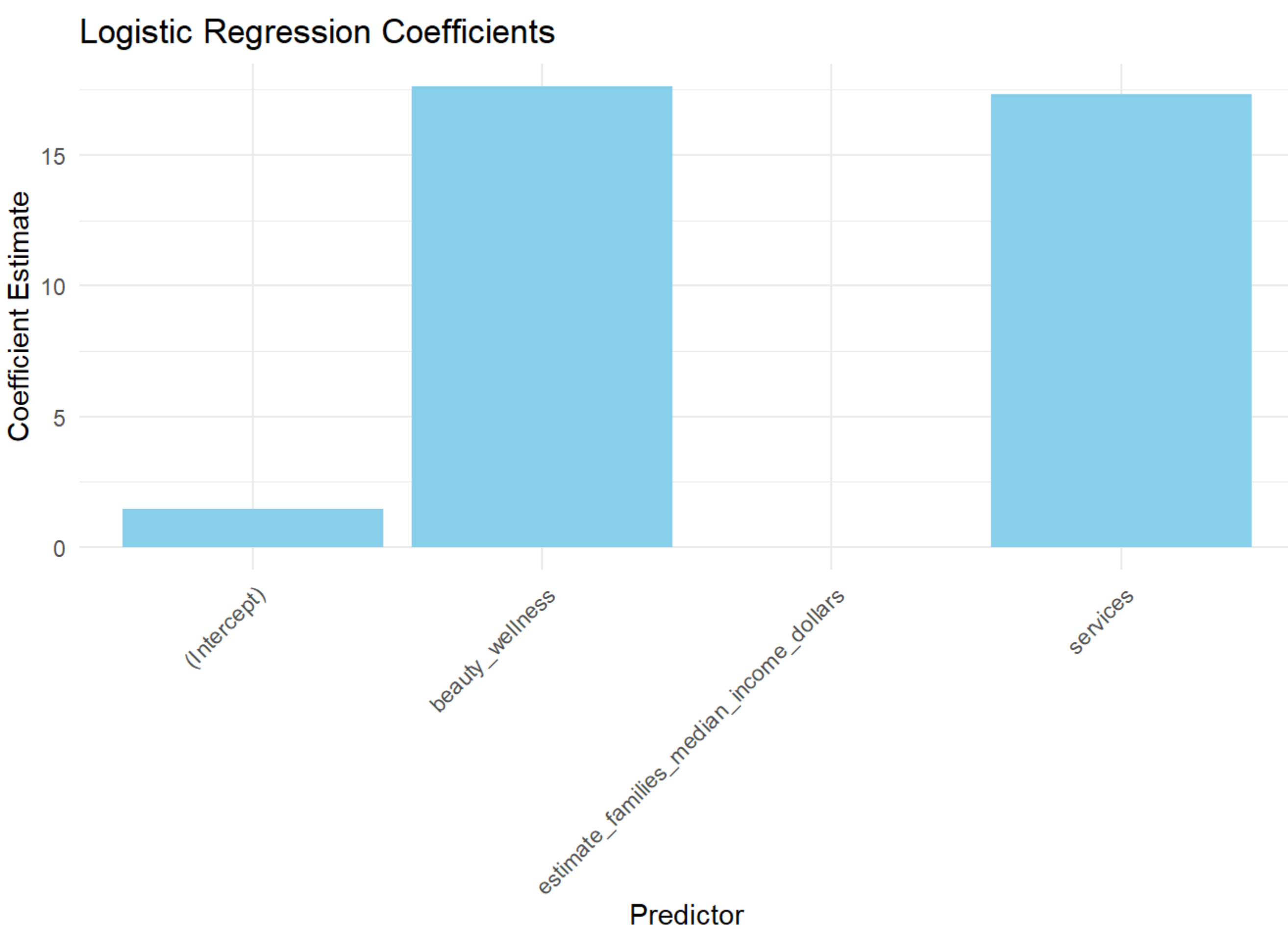
```
## Start:  AIC=73.39
## minority ~ 1
##
##                                           Df Deviance    AIC
## + estimate_families_median_income_dollars  1   66.596 70.596
## + beauty_wellness                          1   68.994 72.994
## + services                                 1   68.994 72.994
## <none>                                         71.393 73.393
## + distance                                 1   69.891 73.891
## + creative_economy                         1   69.968 73.968
## + food                                     1   70.840 74.840
## + angle                                    1   71.107 75.107
## + education                                1   71.267 75.267
## + retail                                   1   71.303 75.303
## + healthcare                               1   71.303 75.303
## + professional_services                    1   71.312 75.312
## + entertainment_culture                    1   71.366 75.366
##
## Step:  AIC=70.6
## minority ~ estimate_families_median_income_dollars
##
##                         Df Deviance    AIC
## + beauty_wellness        1   63.801 69.801
## + services               1   64.421 70.421
## <none>                       66.596 70.596
## + distance               1   65.634 71.634
## + creative_economy       1   65.800 71.800
## + education              1   66.079 72.079
## + angle                  1   66.148 72.148
## + retail                 1   66.430 72.430
## + entertainment_culture  1   66.439 72.439
## + food                   1   66.473 72.473
## + professional_services  1   66.565 72.565
## + healthcare             1   66.594 72.594
##
## Step:  AIC=69.8
## minority ~ estimate_families_median_income_dollars + beauty_wellness
##
##                         Df Deviance    AIC
## + services               1   61.497 69.497
## <none>                       63.801 69.801
## + distance               1   62.543 70.543
## + creative_economy       1   62.879 70.879
## + education              1   63.409 71.409
## + entertainment_culture  1   63.549 71.549
## + angle                  1   63.555 71.555
## + retail                 1   63.663 71.663
## + food                   1   63.748 71.748
## + healthcare             1   63.797 71.797
## + professional_services  1   63.800 71.800
##
```

```
## Step:  AIC=69.5
## minority ~ estimate_families_median_income_dollars + beauty_wellness +
##     services
##
##                          Df Deviance    AIC
## <none>                        61.497 69.497
## + creative_economy        1   60.396 70.396
## + distance                1   60.638 70.638
## + entertainment_culture   1   61.195 71.195
## + education               1   61.246 71.246
## + retail                  1   61.254 71.254
## + angle                   1   61.271 71.271
## + professional_services   1   61.449 71.449
## + healthcare              1   61.458 71.458
## + food                    1   61.475 71.475
```

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

```
##
## Call:
## glm(formula = minority ~ estimate_families_median_income_dollars +
##     beauty_wellness + services, family = binomial, data = women_income_knn)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             1.444e+00  7.080e-01   2.039   0.0414
## estimate_families_median_income_dollars -9.098e-06  4.257e-06  -2.137   0.0326
## beauty_wellness                         1.762e+01  2.484e+03   0.007   0.9943
## services                                1.733e+01  2.716e+03   0.006   0.9949
##
## (Intercept)                             *
## estimate_families_median_income_dollars *
## beauty_wellness
## services
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 71.393  on 51  degrees of freedom
## Residual deviance: 61.497  on 48  degrees of freedom
## AIC: 69.497
##
## Number of Fisher Scoring iterations: 16
```

```
# 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 Coefficient
s") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels by 45 degre
es
```



```
# 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 13  7
##               1 10 22
```

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

```
## [1] 0.6730769
```