

# Editing Data

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```
# Load necessary libraries
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v readr      2.1.4
```

```
## v forcats    1.0.0      v stringr    1.5.0
```

```
## v ggplot2    3.4.3      v tibble     3.2.1
```

```
## v lubridate  1.9.3      v tidyr      1.3.0
```

```
## v purrr      1.0.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(psych)
```

```
##
```

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

```
##
```

```
## The following objects are masked from 'package:ggplot2':
```

```
##
```

```
##      %+%, alpha
```

```
library(stats)
```

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
```

```
## Attaching package: 'Matrix'
```

```
##
```

```
## The following objects are masked from 'package:tidyr':
```

```
##
```

```
##      expand, pack, unpack
```

```
library(scales)
```

```
##
```

```
## Attaching package: 'scales'
##
## The following objects are masked from 'package:psych':
##
##     alpha, rescale
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(jtools)
library(broom)
```

```
## Registered S3 methods overwritten by 'broom':
##   method      from
##   tidy.glht    jtools
##   tidy.summary.glht jtools
```

```
loading_zipdata <- "~/repos/Diversity-Richness/Zip.Code.Datasets/zip_data_unedited_nolabels.csv"
#making a pathway to the downloaded census data
Unedited_zipdata <- read.csv(loading_zipdata) #loading the census data
```

```
reduced_collumns_zip <- c(1:2, 33:34, 42, 47:146) #removing unnecessary columns
```

```
Zipcode_Census_data<- Unedited_zipdata[ , reduced_collumns_zip]
#make a new table without unnecessary columns
```

```
colnames(Zipcode_Census_data) <- c("FIPS", "Name", "3_Digit_Tabulation", "5_Digit_Tabulation" , "Area_Names")
#writing usable column names
```

```
write.csv(Zipcode_Census_data, file = "~/repos/Diversity-Richness/Zip.Code.Datasets/Zipcode_Census_data.csv",
          row.names = FALSE)
#saving the table as a csv file
```

```
library(tidyverse)
library(ggplot2)
```

```
ggplot(Zipcode_Census_data) +
  geom_smooth(aes(x = X_Total_Population_White_Alone, y = Median_Household_Income, color = "#A1C9F4"),
  geom_smooth(aes(x = X_Total_Population_Black_or_African_American_Alone, y = Median_Household_Income, color = "#8DE5A1"),
  geom_smooth(aes(x = X_Total_Population_Asian_Alone, y = Median_Household_Income, color = "#FF9F9B"), fill = "#FF9F9B", alpha = 0.1),
  geom_smooth(aes(x = X_Total_Population_American_Indian_And_Native_Alaskan_Alone, y = Median_Household_Income, color = "#D0BBFF"), fill = "#D0BBFF", alpha = 0.1),
  geom_smooth(aes(x = X_Hispanic, y = Median_Household_Income, color = "#A1C9F4"), fill = "#A1C9F4", alpha = 0.1),
  scale_color_manual(values = c( "#8DE5A1" = "#8DE5A1", "#A1C9F4" = "#A1C9F4", "#D0BBFF" = "#D0BBFF", "#FF9F9B" = "#FF9F9B"),
    name = "Race",
    labels = c("Asian", "White", "Native American/Alaskan", "Hispanic", "Black")) +
  labs(x = "Percentage Population of Population for Each Racial Demographic", y = "Median Household Income")
```

```
## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').
```

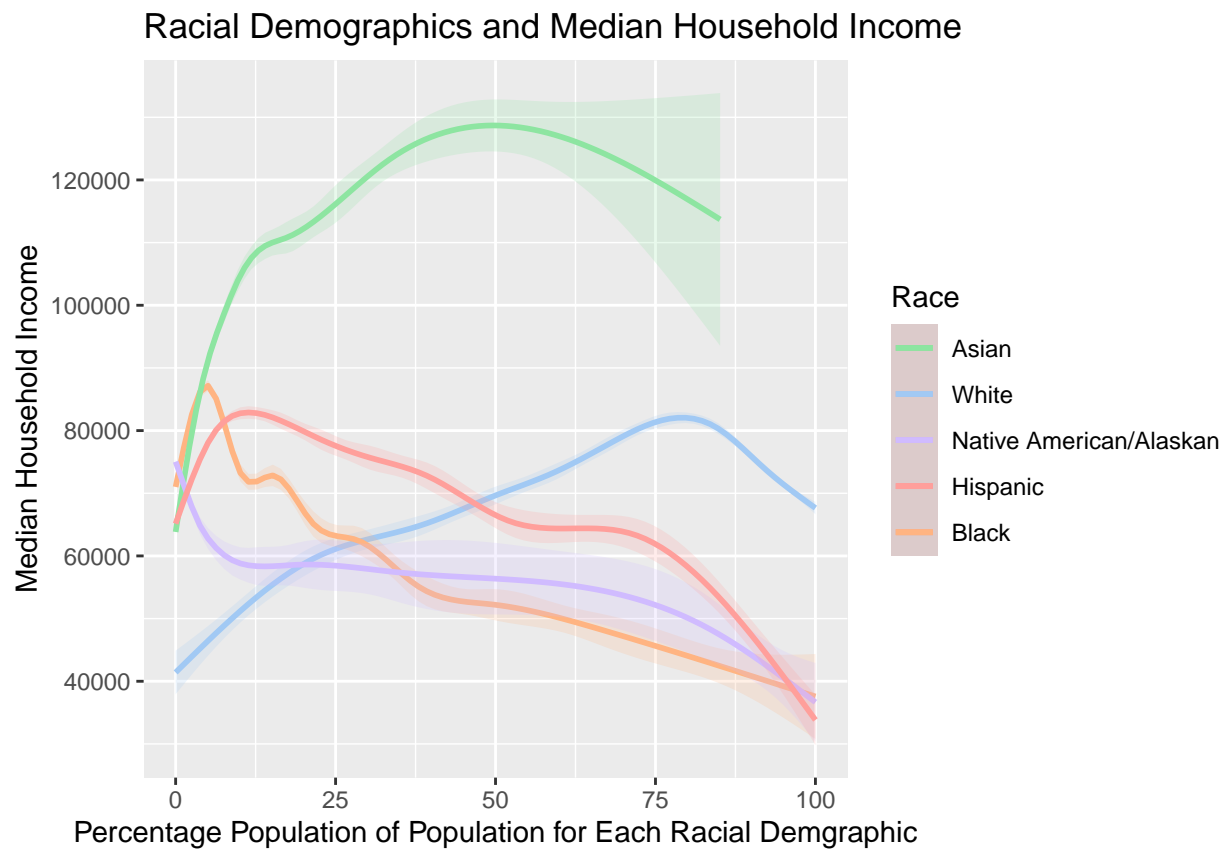


Figure 1: Race-Income Plot

*#Don't know how to correct red shading on key*

```
library(tidyverse)
library(ggplot2)

ggplot(Zipcode_Census_data) +
  geom_smooth(aes(x = X_Total_Population_White_Alone, y = Median_Household_Income, color = "#A1C9F4"), )
  geom_smooth(aes(x = X_Total_Population_Black_or_African_American_Alone, y = Median_Household_Income, color = "#A1C9F4"), )
  geom_smooth(aes(x = X_Total_Population_Asian_Alone, y = Median_Household_Income, color = "#8DE5A1"), )
  geom_smooth(aes(x = X_Total_Population_American_Indian_And_Native_Alaskan_Alone, y = Median_Household_Income, color = "#8DE5A1"), )
  geom_smooth(aes(x = X_Hispanic, y = Median_Household_Income, color = "#FF9F9B"), fill = "#FF9F9B", alpha = 0.5)
  scale_color_manual(values = c( "#8DE5A1" = "#8DE5A1", "#A1C9F4" = "#A1C9F4", "#D0BBFF" = "#D0BBFF", "#FF9F9B" = "#FF9F9B" ),
    name = "Race",
    labels = c("Asian", "White", "Native American/Alaskan", "Hispanic", "Black")) +
  labs(x = "Percentage Population of Population for Each Racial Demographic", y = "Median Household Income")

## 'geom_smooth()' using formula = 'y ~ x'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using formula = 'y ~ x'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using formula = 'y ~ x'

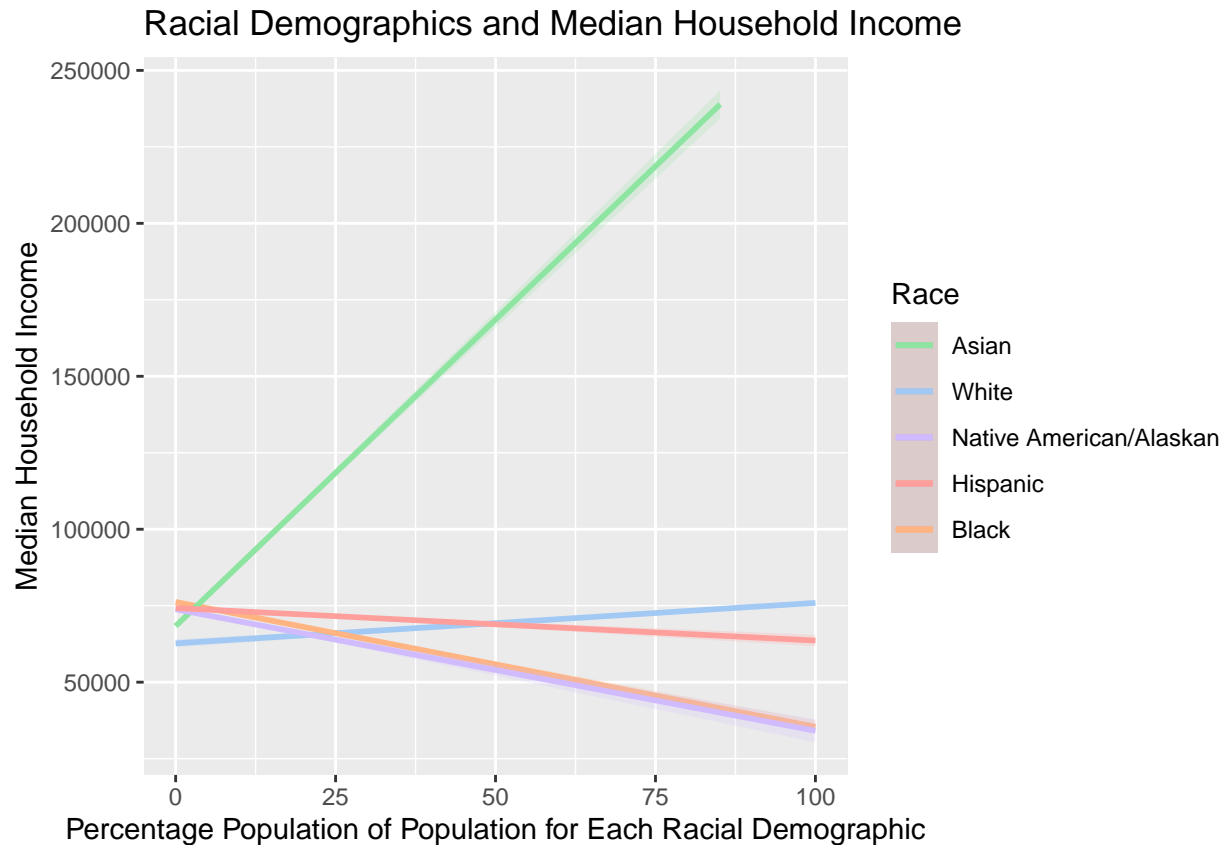
## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using formula = 'y ~ x'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').

## 'geom_smooth()' using formula = 'y ~ x'

## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').
```



```

Zipcode_with_Zones <- Zipcode_Census_data %>%
  mutate(Zipcode_Zone= case_when(
    substr(Name, 7, 8) == "00" ~ "Puerto Rico",
    substr(Name, 7, 7) == "0" ~ "North East (Around MA)",
    substr(Name, 7, 7) == "1" ~ "Lower North East (Arond NY)",
    substr(Name, 7, 7) == "2" ~ "Central East Coast (Around VA)",
    substr(Name, 7, 7) == "3" ~ "South East (Around FL)",
    substr(Name, 7, 7) == "4" ~ "Great Lakes (Around MI)",
    substr(Name, 7, 7) == "5" ~ "Northern Midwest (Around MN)",
    substr(Name, 7, 7) == "6" ~ "Central Interior (Around IL)",
    substr(Name, 7, 7) == "7" ~ "Central South (Around TX)",
    substr(Name, 7, 7) == "8" ~ "Western Interior (Around CO)",
    substr(Name, 7, 7) == "9" ~ "West Coast (includes Hawaii/Alaska)",
    TRUE ~ "Other" # Default case
  ), .after = Area_Name)

Zipcode_with_States <- Zipcode_with_Zones %>%
  mutate(State_Territory= case_when(
    between(as.integer(substr(Name, 7, 9)), 039, 049) ~ "ME",
    between(as.integer(substr(Name, 7, 9)), 030, 038) ~ "NH",
    between(as.integer(substr(Name, 7, 9)), 010, 027) ~ "MA",
    between(as.integer(substr(Name, 7, 9)), 028, 029) ~ "RI",
    between(as.integer(substr(Name, 7, 9)), 150, 196) ~ "PA",
    between(as.integer(substr(Name, 7, 9)), 197, 199) ~ "DE",
    between(as.integer(substr(Name, 7, 9)), 206, 219) ~ "MD",
    between(as.integer(substr(Name, 7, 9)), 200, 205) ~ "DC",
  )

```

```

between(as.integer(substr(Name, 7, 9)), 220, 246) ~ "VA",
between(as.integer(substr(Name, 7, 9)), 247, 269) ~ "WV",
between(as.integer(substr(Name, 7, 9)), 386, 399) ~ "MS",
between(as.integer(substr(Name, 7, 9)), 370, 385) ~ "TN",
between(as.integer(substr(Name, 7, 9)), 700, 715) ~ "LA",
between(as.integer(substr(Name, 7, 9)), 716, 729) ~ "AR",
between(as.integer(substr(Name, 7, 9)), 550, 567) ~ "MN",
between(as.integer(substr(Name, 7, 9)), 820, 831) ~ "WY",
between(as.integer(substr(Name, 7, 9)), 832, 839) ~ "ID",
between(as.integer(substr(Name, 7, 9)), 870, 884) ~ "NM",
between(as.integer(substr(Name, 7, 9)), 889, 899) ~ "NV",
between(as.integer(substr(Name, 7, 9)), 900, 961) ~ "CA",
between(as.integer(substr(Name, 7, 9)), 980, 994) ~ "WA",
between(as.integer(substr(Name, 7, 9)), 967, 968) ~ "HI",
between(as.integer(substr(Name, 7, 9)), 995, 999) ~ "AK",
between(as.integer(substr(Name, 7, 9)), 962, 966) ~ "AP",
between(as.integer(substr(Name, 7, 9)), 006, 009) ~ "PR/VI",
between(as.integer(substr(Name, 7, 8)), 10, 14) ~ "NY",
between(as.integer(substr(Name, 7, 8)), 07, 08) ~ "NJ",
between(as.integer(substr(Name, 7, 8)), 27, 28) ~ "NC",
between(as.integer(substr(Name, 7, 8)), 30, 31) ~ "GA",
between(as.integer(substr(Name, 7, 8)), 32, 34) ~ "FL",
between(as.integer(substr(Name, 7, 8)), 35, 36) ~ "AL",
between(as.integer(substr(Name, 7, 8)), 40, 42) ~ "KY",
between(as.integer(substr(Name, 7, 8)), 43, 45) ~ "OH",
between(as.integer(substr(Name, 7, 8)), 46, 47) ~ "IN",
between(as.integer(substr(Name, 7, 8)), 48, 49) ~ "MI",
between(as.integer(substr(Name, 7, 8)), 50, 12) ~ "IA",
between(as.integer(substr(Name, 7, 8)), 53, 54) ~ "WI",
between(as.integer(substr(Name, 7, 8)), 60, 62) ~ "IL",
between(as.integer(substr(Name, 7, 8)), 63, 65) ~ "MO",
between(as.integer(substr(Name, 7, 8)), 66, 67) ~ "KS",
between(as.integer(substr(Name, 7, 8)), 68, 69) ~ "NE",
between(as.integer(substr(Name, 7, 8)), 73, 74) ~ "OK",
between(as.integer(substr(Name, 7, 8)), 75, 79) ~ "TX",
between(as.integer(substr(Name, 7, 8)), 80, 81) ~ "CO",
between(as.integer(substr(Name, 7, 8)), 85, 86) ~ "AZ",
str_detect(Name, "05") ~ "VT",
str_detect(Name, "06") ~ "CT",
str_detect(Name, "29") ~ "SC",
str_detect(Name, "57") ~ "SD",
str_detect(Name, "58") ~ "ND",
str_detect(Name, "59") ~ "MT",
str_detect(Name, "84") ~ "UT",
str_detect(Name, "97") ~ "OR",
str_detect(Name, "09") ~ "AE",
str_detect(Name, "340") ~ "AA",
str_detect(Name, "969") ~ "PW/FM/MH/MP/GU",
str_detect(Name, "96799") ~ "AS",
TRUE ~ "Other" # Default case
), .after = Zipcode_Zone)

```

```
write.csv(Zipcode_with_States, file = "~/repos/Diversity-Richness/Zip.Code.Datasets/Zipcode_data_with_A")
```

```
# Create the boxplot
ggplot(Zipcode_with_Zones, aes(x = Zipcode_Zone, y = X_Total_Population_White_Alone, fill = Zipcode_Zone)) +
  geom_boxplot(outlier.shape = NA) +
  labs(x = "ZIP Code Zones", y = "Percentage White",
       title = "Percentage of Population White by Zip Code Region", fill = "Zone Names") +
  theme(axis.text.x = element_blank())
```

## Warning: Removed 587 rows containing non-finite values ('stat\_boxplot()').

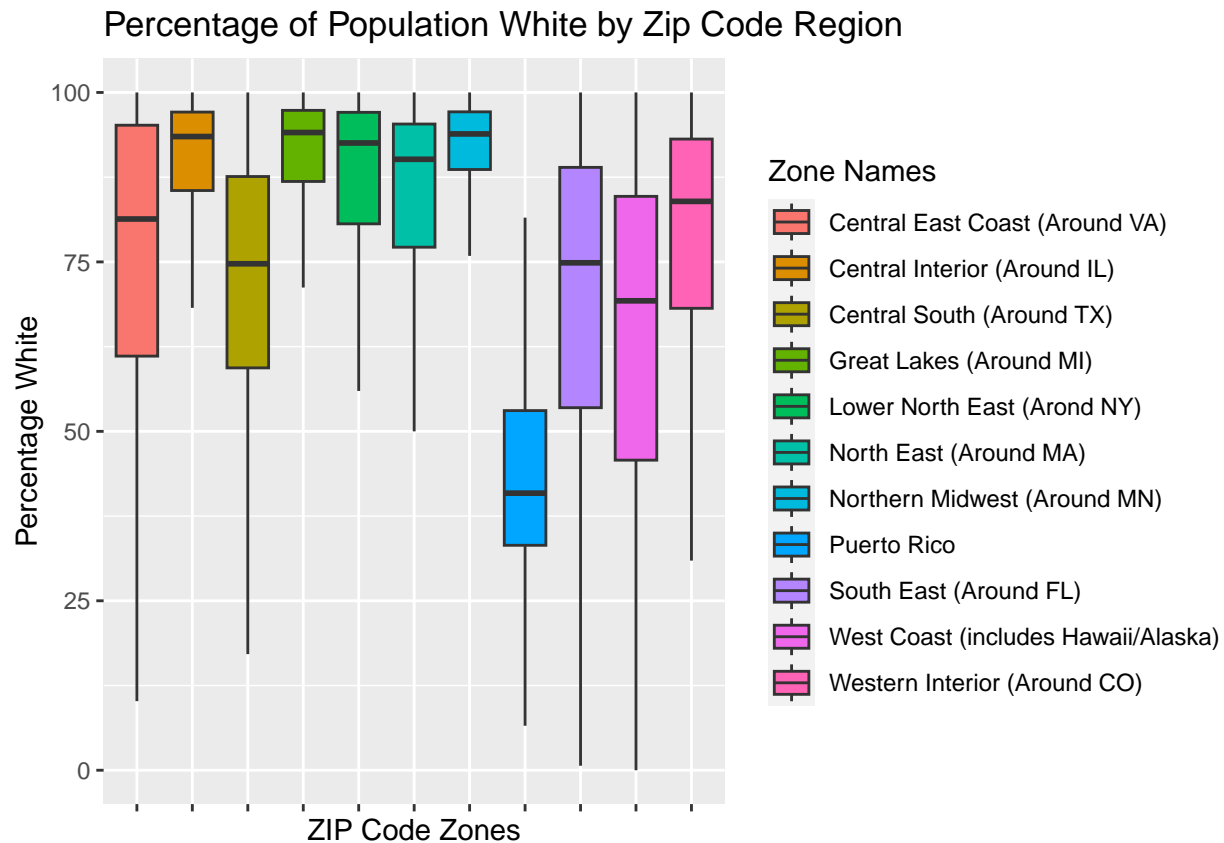


Figure 2: (ref:Whiteness-Income-Boxplot-Caption)

(ref:ZIP-Region-Table-Caption) ZIP Region Table.

```
Zipcode_with_States %>%
  group_by(Zipcode_Zone) %>%
  summarize(
    mean.percent.white = mean(X_Total_Population_White_Alone, na.rm = T),
    sd.percent.white = sd(X_Total_Population_White_Alone, na.rm = T),
    mean.percent.black = mean(X_Total_Population_Black_or_African_American_Alone, na.rm = T),
    sd.percent.black = sd(X_Total_Population_Black_or_African_American_Alone, na.rm = T),
    mean.percent.asian = mean(X_Total_Population_Asian_Alone, na.rm = T),
    sd.percent.asian = sd(X_Total_Population_Asian_Alone, na.rm = T),
    mean.percent.native.american = mean(X_Total_Population_American_Indian_And_Native_Alaskan_Alone, na.rm = T),
    sd.percent.native.american = sd(X_Total_Population_American_Indian_And_Native_Alaskan_Alone, na.rm = T)
```

Table 1: Average ZIP Code Demographics by Region

Region	% White	SD White	% Black	SD Black	% Asian	SD Asian	% Native American	SD Native American	% Hispanic	SD Hispanic	ZIP Codes Per Region
Central East Coast (Around VA)	<b>75.02</b>	23.79	<b>15.97</b>	20.58	<b>1.85</b>	4.47	<b>0.53</b>	3.80	<b>5.47</b>	7.66	<b>3452</b>
Central Interior (Around IL)	<b>87.65</b>	16.50	<b>4.12</b>	12.27	<b>1.34</b>	3.58	<b>0.58</b>	3.36	<b>5.78</b>	9.86	<b>3721</b>
Central South (Around TX)	<b>71.48</b>	20.73	<b>11.12</b>	17.71	<b>1.62</b>	3.89	<b>2.05</b>	5.32	<b>19.83</b>	24.46	<b>3808</b>
Great Lakes (Around MI)	<b>88.36</b>	16.24	<b>5.32</b>	13.45	<b>1.05</b>	2.67	<b>0.32</b>	1.56	<b>3.47</b>	5.70	<b>3812</b>
Lower North East (Arond NY)	<b>84.84</b>	19.50	<b>5.30</b>	12.05	<b>2.59</b>	6.03	<b>0.29</b>	2.54	<b>6.36</b>	10.47	<b>3726</b>
North East (Around MA)	<b>83.01</b>	18.66	<b>4.39</b>	9.62	<b>3.95</b>	7.19	<b>0.32</b>	1.56	<b>8.06</b>	12.45	<b>2445</b>
Northern Midwest (Around MN)	<b>89.37</b>	15.50	<b>1.48</b>	5.30	<b>1.05</b>	2.77	<b>3.08</b>	12.65	<b>3.74</b>	5.62	<b>3766</b>
Puerto Rico	<b>43.15</b>	18.15	<b>9.12</b>	9.05	<b>0.23</b>	0.45	<b>0.15</b>	0.26	<b>97.96</b>	6.07	<b>132</b>
South East (Around FL)	<b>68.73</b>	24.74	<b>21.41</b>	24.39	<b>1.55</b>	3.32	<b>0.33</b>	1.23	<b>9.14</b>	14.17	<b>3483</b>
West Coast (includes Hawaii/Alaska)	<b>63.64</b>	26.01	<b>3.04</b>	5.78	<b>7.91</b>	12.41	<b>5.51</b>	17.96	<b>22.06</b>	23.07	<b>3177</b>
Western Interior (Around CO)	<b>76.22</b>	24.31	<b>1.77</b>	4.17	<b>1.43</b>	2.87	<b>6.29</b>	20.06	<b>20.39</b>	23.04	<b>2252</b>

```

mean.percent.hispanic = mean(X_Hispanic, na.rm = T),
sd.percent.hispanic = sd(X_Hispanic, na.rm = T),
n.per.region = n()
) %>%
knitr::kable("latex", col.names = c("Region", "% White", "SD White", "% Black", "SD Black", "% Asian",
  digits = 2,
  align = "lrlrlrlrlrlr",
  caption = "Average ZIP Code Demographics by Region") %>%
kableExtra::kable_styling(position = "left", latex_options = "scale_down") %>%
kableExtra::column_spec(seq(2, 12, 2), bold = TRUE)

```

```

#for some reason scale down is not working
#struggling with captioning and referencing

```

I will now refer to “Figure @ref(fig:Smooth-Plot-Income-Race)”, Figure @ref(fig:Smooth-Plot-Income-Race), Figure @ref(fig:Race-Income Plot), Figure @ref(fig:Race-Income Plot)

Descriptive Chunk:

```
library(papaja)
```

```
## Loading required package: tinylabls
```

```
##
```

```
## Attaching package: 'papaja'
```

```
## The following object is masked from 'package:jtools':
```

```
##
```

```
## theme_ap
```

```
Census.desc <- Zipcode_with_States %>%
```

```
  select(X_Total_Population_White_Alone, Median_Household_Income) %>%
  drop_na()
```

```
Census.desc.long <- Census.desc %>%
```

```
  pivot_longer(c(X_Total_Population_White_Alone, Median_Household_Income), names_to = "measure")
```

```
Census.desc.long %>%
```

```
  group_by(measure) %>%
  summarize(mean = mean(value),
    median = median(value),
```



Table 2:

measure	mean	median	sd	first_quartile	third_quartile	range
Median_Household_Income	73,132.63	66,993.00	31,359.58	53,466.00	85,313.00	247,502.00
X_Total_Population_White_Alone	78.92	87.45	22.12	69.49	95.04	100.00

```

sd = sd(value),
first_quartile = quantile(value, probs = c(.25)),
third_quartile = quantile(value, probs = c(.75)),
range = diff(range(value))) %>%
apa_table()

```

Hypothesis Testing: Whiteness/Income

```

whiteness_income_corr <- cor(Census.desc)
(whiteness_income_corr_simple <- cor(Census.desc[,1], Census.desc[,2]))

```

```
## [1] 0.09329494
```

```

whiteness_income_corr2 <- corr.test(Census.desc) #run corr
whiteness_income_corr2$p.adj

```

```
## [1] 3.656746e-60
```

```

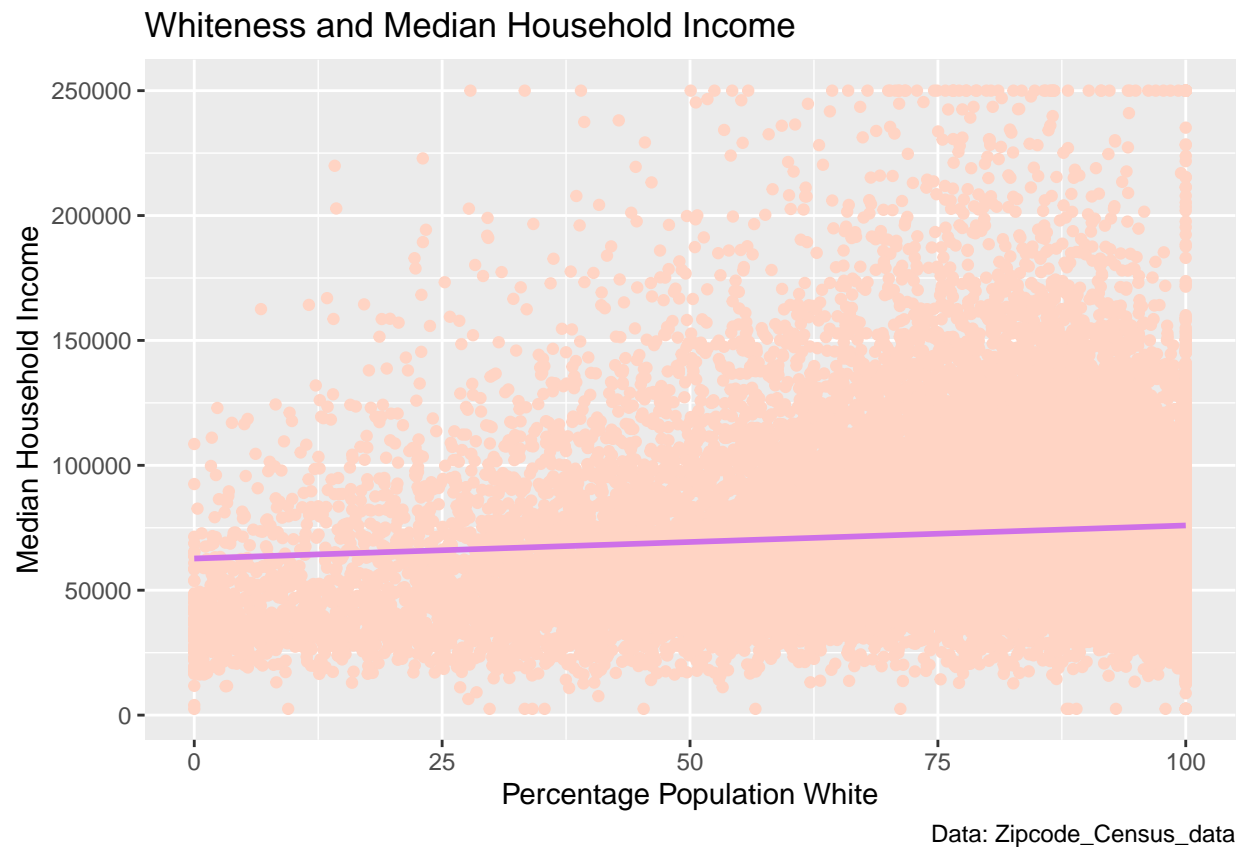
ggplot(Zipcode_Census_data) +
  geom_jitter(aes(x = X_Total_Population_White_Alone, y = Median_Household_Income), fill = "black", color = "black") +
  geom_smooth(aes(x = X_Total_Population_White_Alone, y = Median_Household_Income), fill = "#ffb5e2", color = "black") +
  labs(x = "Percentage Population White", y = "Median Household Income",
       title = "Whiteness and Median Household Income", caption = "Data: Zipcode_Census_data")

```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

```
## Warning: Removed 3153 rows containing non-finite values ('stat_smooth()').
```

```
## Warning: Removed 3153 rows containing missing values ('geom_point()').
```



```
corrplot::corrplot(whiteness_income_corr)
```



```
Census_desc2 <- Zipcode_with_States %>%
  select("X_Total_Population_White_Alone", "X_Total_Population_Black_or_African_American_Alone", "X_Tot
  drop_na()
```

```
less_diversity_ttest <- t.test(
  filter(Census_desc2, X_Total_Population_White_Alone >= 75)$Median_Household_Income,
  filter(Census_desc2, X_Total_Population_White_Alone <= 75)$Median_Household_Income)
```

```
less_diversity_ttest
```

```
##
## Welch Two Sample t-test
##
## data: filter(Census_desc2, X_Total_Population_White_Alone >= 75)$Median_Household_Income and filter
## t = 10.021, df = 15398, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 3348.951 4977.716
## sample estimates:
## mean of x mean of y
## 74406.65 70243.32
```

```
asian_westcoast_test <- t.test(
  filter(Census_desc2, X_Total_Population_Asian_Alone >= 50, Zipcode_Zone == "West Coast (includes
```

```

    filter(Census_desc2, X_Total_Population_White_Alone <= 50, Zipcode_Zone == "West Coast (includes
asian_westcoast_test

```

```

##
## Welch Two Sample t-test
##
## data: filter(Census_desc2, X_Total_Population_Asian_Alone >= 50, Zipcode_Zone == "West Coast (includes
## t = 7.547, df = 70.625, p-value = 1.187e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 35308.15 60667.43
## sample estimates:
## mean of x mean of y
## 130621.3 82633.5

```

```

(income_race_model <- lm(data = Census_desc2, Median_Household_Income ~ X_Total_Population_White_Alone))

```

```

##
## Call:
## lm(formula = Median_Household_Income ~ X_Total_Population_White_Alone,
##     data = Census_desc2)
##
## Coefficients:
##              (Intercept)  X_Total_Population_White_Alone
##                62693.0                132.3

```

```

(income_race_model_sum <- summary(income_race_model))

```

```

##
## Call:
## lm(formula = Median_Household_Income ~ X_Total_Population_White_Alone,
##     data = Census_desc2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -73423 -19799  -6711  11852 183624
##
## Coefficients:
##              (Intercept)      X_Total_Population_White_Alone
##                62693.039                132.286
##
## Estimate Std. Error t value Pr(>|t|)
## (Intercept)    62693.039     661.225   94.81  <2e-16 ***
## X_Total_Population_White_Alone    132.286      8.068   16.40  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31220 on 30619 degrees of freedom
## Multiple R-squared:  0.008704, Adjusted R-squared:  0.008672
## F-statistic: 268.8 on 1 and 30619 DF, p-value: < 2.2e-16

```

```

income_race_tidy <- tidy(income_race_model)

```

In Model 1, which includes just one independent variable (`Median_Household_Income ~ X_Total_Population_White_Alone`) the percentage of the ZIP Code population that is white is positively associated with median household income ( $\beta = 132.286$ ,  $p < 3.6567462 \times 10^{-60}$ ). The Intercept, or approximated mean of median household income if the percent white were zero, is \$ 62693.04 , and for every additional unit of percent of the population white, we expect an increase of \$ 132.286. (Additionally, it is important to note that the Intercept is different than the mean median household income for all ZIP Codes \$ 73132.63 .) Therefore, if we increased percent white by one unit that manner, we would expect the mean ZIP Code median household income to increase to \$ 62825.33 . Additionally, Model 1 shows a very significant relationship between percent of the population white and median household income ( $< .001$ ) indicating a strong association between the two variables.

$$6.2693039 \times 10^4$$