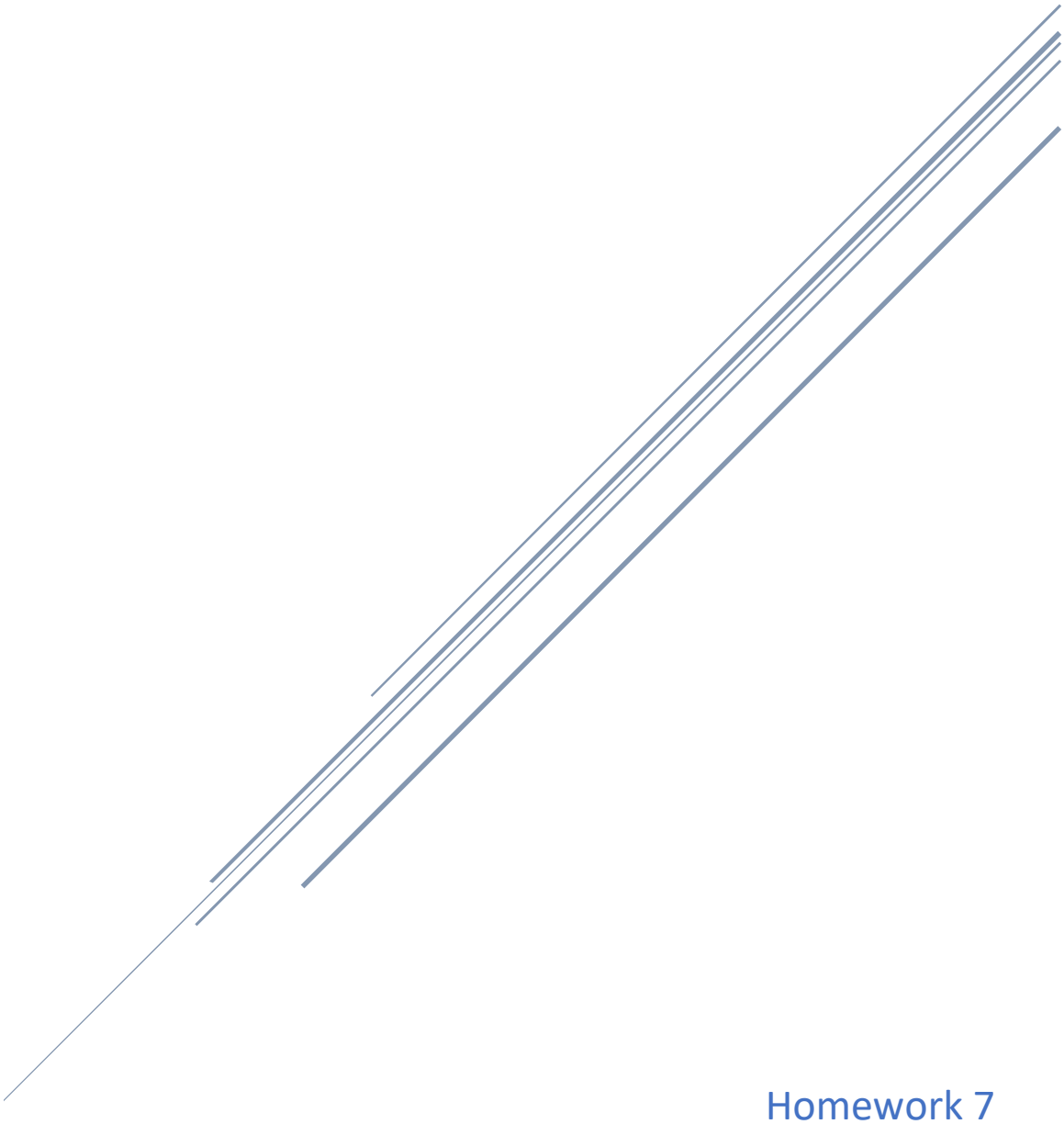
IST 687

Marriah Lewis



Homework 7
Due Week 7 and 11/22/2020

```
#Install read xlsx package
>install.packages("readxl")
>library(readxl)
# Load the data
>income_df<- read_excel("C:\\Users\\lewis\\data_science\\MedianZIP_3.xlsx")
>View(income_df)
#Clean up the df; rename the column names
>colnames(income_df)<-c("zip", "median", "mean", "population")
>str(income_df)
tibble [32,634 x 4] (S3: tbl_df/tbl/data.frame)
 $ zip       : num [1:32634] 1001 1002 1003 1005 1007 ...
 $ median    : num [1:32634] 56663 49853 28462 75423 79076 ...
 $ mean      : num [1:32634] 66688 75063 35121 82442 85802 ...
 $ population: num [1:32634] 16445 28069 8491 4798 12962 ...
>summary(income_df)
   zip          median           mean          population
 Min.   : 1001   Min.   :    32.98   Min.   :    53.6   Min.   :    1
 1st Qu.:27301   1st Qu.: 38462.00   1st Qu.: 48593.2   1st Qu.:   736
 Median :49875   Median : 46503.32   Median : 56949.6   Median :  2756
 Mean   :49875   Mean   : 50938.21   Mean   : 63452.2   Mean   :  9193
 3rd Qu.:72134   3rd Qu.: 58255.50   3rd Qu.: 70341.2   3rd Qu.: 12513
 Max.   :99929   Max.   :223106.17   Max.   :361842.3   Max.   :113916
                                                        NA's   :7
#Function to remove NA's
>remove_na<- function(df, n=0){
  df[rowSums(is.na(df)) <=n,]
}
#Remove NA's in income_df
>income_df<- remove_na(income_df)
```

```
>summary(income_df)
zip            median           mean          population
 Min.   : 1001   Min.   :    32.98   Min.   :    53.6   Min.   :    1
 1st Qu.:27300   1st Qu.: 38462.94   1st Qu.: 48593.2   1st Qu.:   736
 Median :49872   Median : 46502.73   Median : 56949.6   Median :  2757
 Mean   :49870   Mean   : 50938.83   Mean   : 63452.2   Mean   :  9194
 3rd Qu.:72129   3rd Qu.: 58256.00   3rd Qu.: 70341.2   3rd Qu.: 12516
 Max.   :99929   Max.   :223106.17   Max.   :361842.3   Max.   :113916
#Load zipcode package
>install.packages("zipcodeR")
>library(zipcodeR)
>data("zip_code_db")
>dfZip<-data.frame(zip_code_db)
>head(dfZip)
zipcode zipcode_type     major_city   post_office_city
1   35004     Standard         Moody         Moody, AL
2   35005     Standard     Adamsville     Adamsville, AL
3   35006     Standard          Adger          Adger, AL
4   35007     Standard      Alabaster      Alabaster, AL
5   35010     Standard Alexander City Alexander City, AL
6   35011      PO Box Alexander City             <NA>
                                                                common_city_list
1                 78, 9c, 8b, 56, f2, cd, cf, 4f, a9, 54, d2, 51, 50, 72, 4c, ce, 4d, 2c, 52, 8a, 05, 00, 33, f0, 05, 79
2                       78, 9c, 8b, 56, 72, 4c, 49, cc, 2d, 2e, cb, cc, c9, 49, 55, 8a, 05, 00, 24, 9d, 04, ff
3                             78, 9c, 8b, 56, 72, 4c, 49, 4f, 2d, 52, 8a, 05, 00, 0d, f9, 02, e0
4                           78, 9c, 8b, 56, 72, cc, 49, 4c, 4a, 2c, 2e, 49, 2d, 52, 8a, 05, 00, 1e, e8, 04, 8c
5 78, 9c, 8b, 56, 72, cc, 49, ad, 48, cc, 4b, 49, 2d, 52, 70, ce, 2c, a9, 54, d2, 51, 00, 8b, 40, 38, b1, 00, a4,
59, 0a, 1d
```

6 78, 9c, 8b, 56, 72, cc, 49, ad, 48, cc, 4b, 49, 2d, 52, 70, ce, 2c, a9, 54, d2, 51, 00, 8b, 40, 38, b1, 00, a4, 59, 0a, 1d

	county	state	lat	lng	timezone	radius_in_miles
1	St. Clair County	AL	33.62	-86.50	Central	4
2	Jefferson County	AL	33.59	-86.99	Central	6
3	Jefferson County	AL	33.40	-87.20	Central	11
4	Shelby County	AL	33.22	-86.79	Central	5
5	Tallapoosa County	AL	32.90	-85.90	Central	17
6	Tallapoosa County	AL	NA	NA	<NA>	NA

	area_code_list	population
1	78, 9c, 8b, 56, 32, 32, 30, 55, 8a, 05, 00, 06, 4a, 01, 94	10427
2	78, 9c, 8b, 56, 32, 32, 30, 55, 8a, 05, 00, 06, 4a, 01, 94	7942
3	78, 9c, 8b, 56, 32, 32, 30, 55, 8a, 05, 00, 06, 4a, 01, 94	3121
4	78, 9c, 8b, 56, 32, 32, 30, 55, 8a, 05, 00, 06, 4a, 01, 94	26225
5	78, 9c, 8b, 56, 32, 32, 35, 53, 8a, 05, 00, 06, 61, 01, 9a	20816
6	78, 9c, 8b, 56, 32, 32, 35, 53, 8a, 05, 00, 06, 61, 01, 9a	NA

	population_density	land_area_in_sqmi	water_area_in_sqmi	housing_units
1	577	18.07	0.14	4523
2	230	34.51	0.35	3485
3	31	99.81	3.02	1495
4	702	37.38	0.67	9799
5	96	217.59	25.60	10307
6	NA	NA	NA	NA

	occupied_housing_units	median_home_value	median_household_income	bounds_west
1	4214	142500	58832	-86.55178
2	3067	97000	46059	-87.08163
3	1188	95400	51929	-87.34170
4	9180	153900	64299	-86.86177
5	8476	90800	37380	-86.10822

```
6                NA             NA                  NA        NA
 bounds_east bounds_north bounds_south
1   -86.45282    33.66850    33.56269
2   -86.90677    33.63943    33.53390
3   -87.07163    33.55580    33.32744
4   -86.72683    33.27176    33.15020
5   -85.76372    33.10446    32.69872
6          NA          NA          NA
#clean up dfZip; drop specific columns that are not relevant
>dfZip<-dfZip[,-2]
>dfZip<-dfZip[,-c(3:5)]
>dfZip<-dfZip[,-c(6:20)]
>head(dfZip)
zipcode     major_city state   lat    lng
1   35004         Moody    AL 33.62 -86.50
2   35005     Adamsville    AL 33.59 -86.99
3   35006          Adger    AL 33.40 -87.20
4   35007      Alabaster    AL 33.22 -86.79
5   35010 Alexander City    AL 32.90 -85.90
6   35011 Alexander City    AL    NA     NA
#rename columns in dfZip
>colnames(dfZip)<-c("zip", "city", "state", "latitude", "longitude")
#remove the NA's in dfZip
>dfZip<-remove_na(dfZip)
#Merge dataframe
>merged_df<- merge(income_df, dfZip)
>head(merged_df)
zip    median      mean population     city state latitude longitude
1 10001  71244.61 123112.78      17678 New York    NY   40.750   -73.990
```

```
2 10002  30843.96  46258.61      70878 New York    NY   40.720   -73.990
3 10003  89998.53 139331.00      53609 New York    NY   40.730   -73.990
4 10004 110183.69 156682.76       1271 New York    NY   40.700   -74.020
5 10005 115133.29 163762.66       1517 New York    NY   40.705   -74.005
6 10006 111220.00 156776.00        972 New York    NY   40.708   -74.013
#Remove Hawaii and Alaska
>merged_df<-merged_df[which(!(merged_df$state=='HI')),]
>merged_df<-merged_df[which(!(merged_df$state=='AK')),]
#Create a df with full state name and state abbrev.
>dfState<- data.frame(state.abb, state.name)
>colnames(dfState)<-c("state", "statename")
>head(dfState)
state  statename
1    AL    Alabama
2    AK     Alaska
3    AZ    Arizona
4    AR   Arkansas
5    CA California
6    CO   Colorado
#the difference between the merge df and the dfState
>setdiff(merged_df$state, dfState$state)
[1] "DC"
#drop DC
>merged_df<-merged_df[which(!(merged_df$state=="DC")),]
#Add the state abbreviations and the state names as new columns
>df_merge_states<-merge(dfState, merged_df)
>head(df_merge_states)
state statename   zip   median     mean population     city latitude longitude
1    AL   Alabama 35594 36502.85 48734.62       7794 Winfield    34.00    -87.80
```

```
2    AL   Alabama 35601 38537.53 49601.84     34434  Decatur    34.61    -87.01
3    AL   Alabama 35603 54565.09 63214.74     30545  Decatur    34.50    -87.00
4    AL   Alabama 35610 47610.24 61413.90      2252 Anderson    34.94    -87.24
5    AL   Alabama 35611 37503.67 54412.67     25251   Athens    34.80    -87.10
6    AL   Alabama 35613 58471.47 71060.85     20546   Athens    34.80    -86.90
#Step 2
>install.packages("ggmap")
>library(ggmap)
#Lowercase-state names
>df_merge_states$statename<-tolower(df_merge_states$statename)
>head(df_merge_states)
  state statename   zip   median     mean population     city latitude longitude
1    AL   alabama 35594 36502.85 48734.62       7794 Winfield    34.00    -87.80
2    AL   alabama 35601 38537.53 49601.84      34434  Decatur    34.61    -87.01
3    AL   alabama 35603 54565.09 63214.74      30545  Decatur    34.50    -87.00
4    AL   alabama 35610 47610.24 61413.90       2252 Anderson    34.94    -87.24
5    AL   alabama 35611 37503.67 54412.67      25251   Athens    34.80    -87.10
6    AL   alabama 35613 58471.47 71060.85      20546   Athens    34.80    -86.90
#Df the average median income and population for each state
>library(sqldf)
>med_incomeDF<-sqldf('select statename as state, AVG(median) as medincome, sum(population) as
sumpop from df_merge_states group by state')
>head(med_incomeDF)
    state medincome   sumpop
1    alabama  40512.96  4761097
2   arkansas  36986.64  2912218
3    arizona  48094.30  6356003
4 california  62576.20 36925462
5   colorado  56159.21  4979179
```

6 delaware 64298.63 892487

#Show US Map

>install.packages("maps")

>library(maps)

>USA<-map_data("state")

#Show the US map representing the color with the average median income

>install.packages("mapproj")

>library(mapproj)

>m<-ggplot(med_incomeDF, aes(map_id=state))

>m<- m+geom_map(map = USA, aes(fill=medincome))

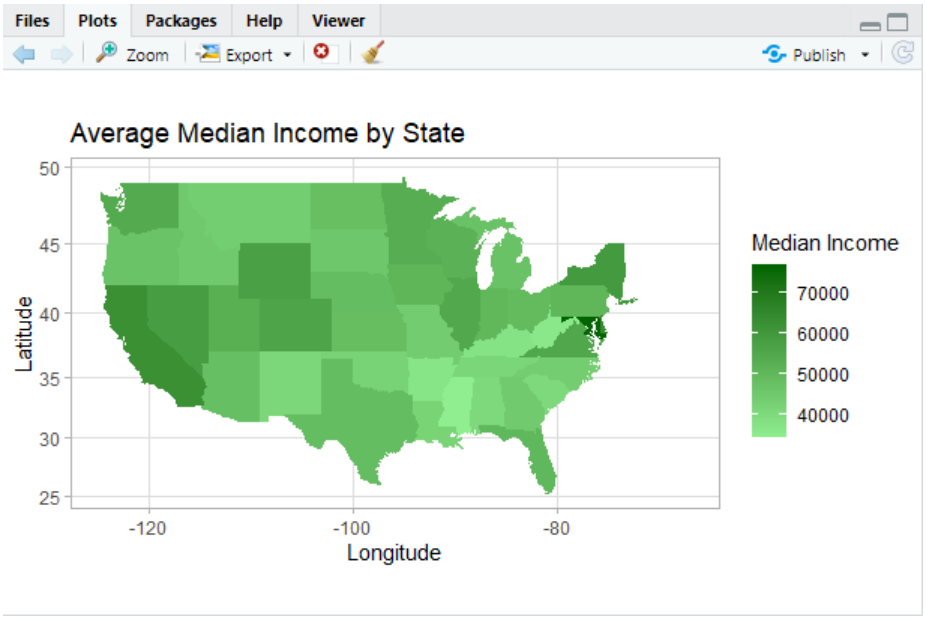
>m<-m+expand_limits(x=USA$long, y=USA$lat)+ coord_map()

>m<-m+ggtitle("Average Median Income by State")+ theme_light()

>m<-m+scale_fill_gradient(low = "lightgreen", high = "darkgreen", name="Median Income")

>m<-m+xlab("Longitude")+ylab("Latitude")

>m



#Map-Population by State

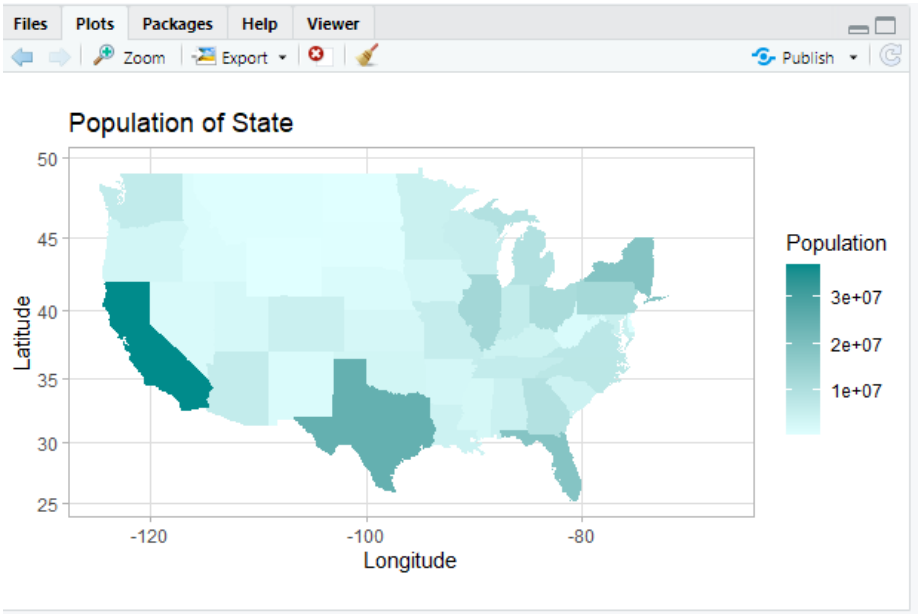
>p<-ggplot(med_incomeDF, aes(map_id=state))+ geom_map(map=USA, aes(fill=med_incomeDF$sumpop))

>p<-p+expand_limits(x=USA$long, y=USA$lat)+ coord_map()

>p<-p+ggtitle("Population of State")+theme_light()

>p<-p+scale_fill_gradient(low="lightcyan", high="darkcyan", name="Population")

>p



#Step 3

#Income per zip code

>i<-ggplot(df_merge_states, aes(map_id=statename))

>i<-i+geom_map(map = USA)

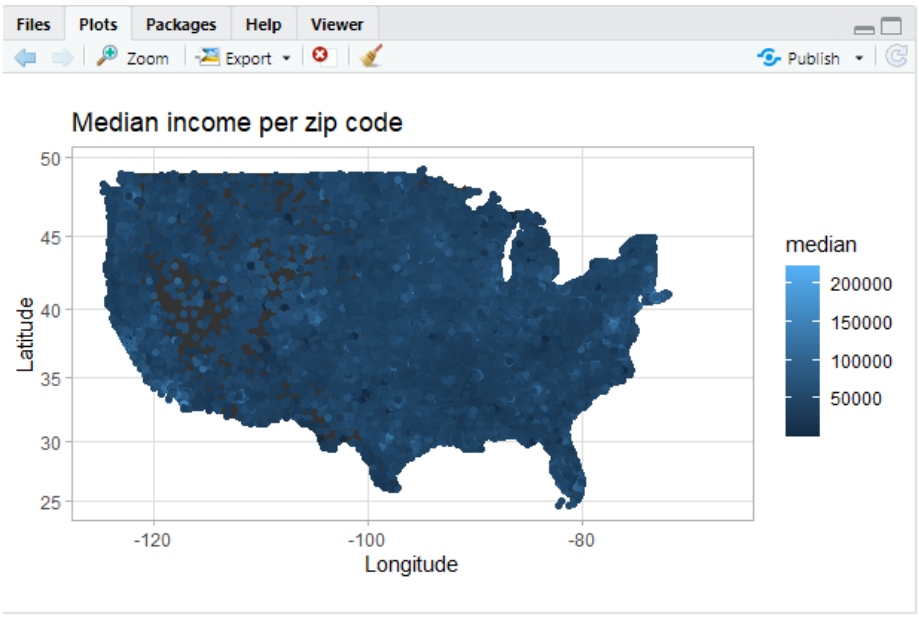
>i<-i+expand_limits(x=USA$long, y=USA$lat)+coord_map()

>i<-i+geom_point(aes(x=df_merge_states$longitude, y=df_merge_states$latitude, color=median))+theme_light()

>i<-i+ggtitle("Median income per zip code")

>i<-i+xlab("Longitude")+ylab("Latitude")

>i



#Step 4

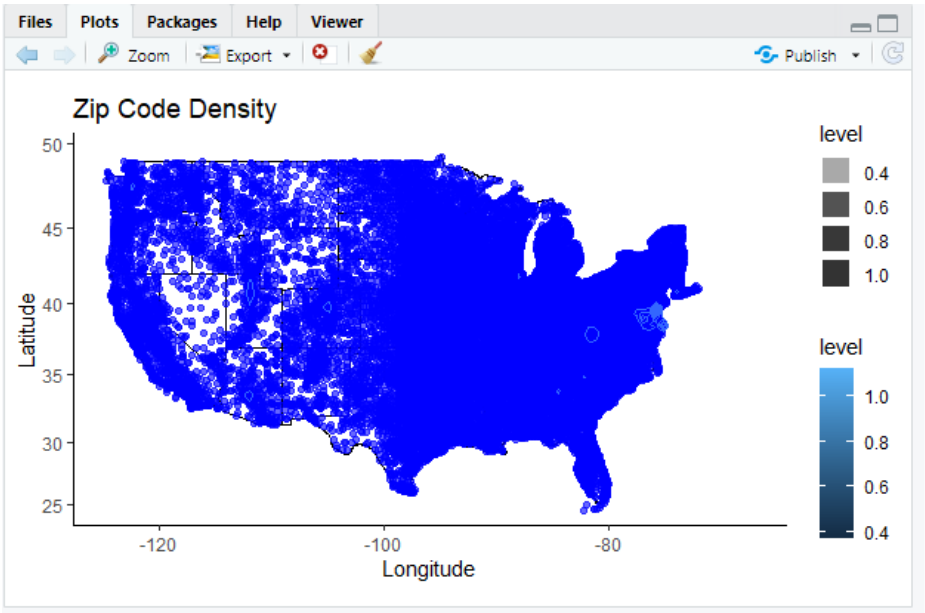
>z<-ggplot(df_merge_states, aes(x=longitude, y=latitude))

>z<-z+geom_point(color="blue", alpha=.6)+stat_density2d()

>z<-z+expand_limits(x=USA$long, y=USA$lat)+coord_map()

>z<-z+theme_classic()+ggtitle("Zip Code Density")

>z<-z+xlab("Longitude")+ylab("Latitude")



#Step 5

#Zoom in to the region around NYC-Income per zip code

> NYC<-ggplot(df_merge_states, aes(map_id=statename))

> NYC<-NYC+geom_map(map=USA)

> NYC <- NYC + xlim(-80,-67) + ylim(37.5,47.5) + coord_map()

> NYC <- NYC + geom_point(aes(x=df_merge_states$longitude, y=df_merge_states$latitude, color= median)) + theme_light()

> NYC <- NYC + ggtitle('Zoomed Median income per zip code')

> NYC <- NYC + xlab('Longitude') + ylab('Latitude')

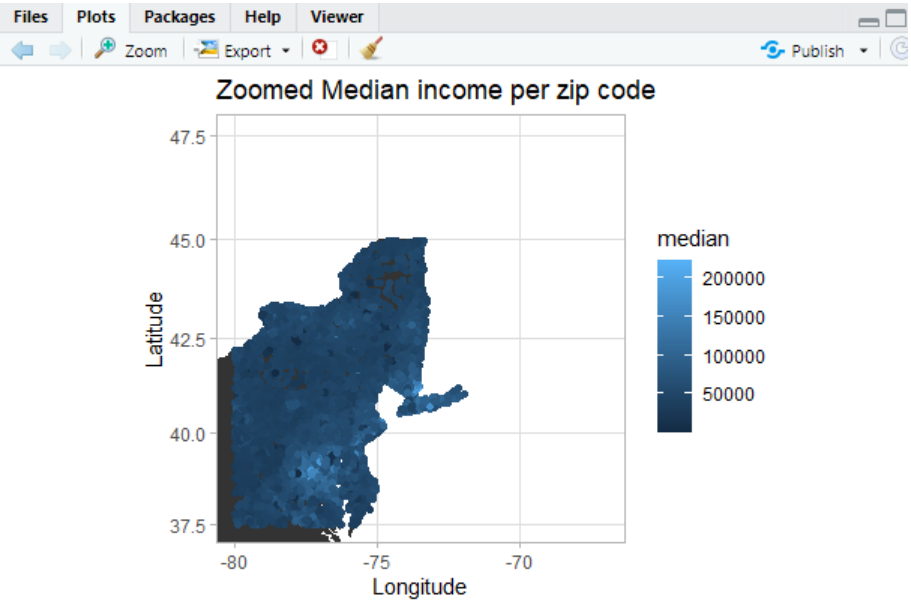
> NYC

Warning messages:

1: Use of `df_merge_states$longitude` is discouraged. Use `longitude` instead.

2: Use of `df_merge_states$latitude` is discouraged. Use `latitude` instead.

3: Removed 25436 rows containing missing values (geom_point).



#Zoom Zip Code density

>z_zoom<-ggplot(df_merged_states, aes(x=longitude, y=latitude))

>z_zoom<-z_zoom+geom_point(color="blue", alpha=.6)+stat_density2d()

>z_zoom<-z_zoom+theme_classic()+ggtitle("Zip Code Density")

```
>z_zoom<-z+xlab("Longitude")+ylab("Latitude")

>z_zoom<-z_zoom+xlim(-80,-67)+ylim(37.5,47.5)+coord_map()
```