Effect of Race, Ethnicity and Gender on Melanoma cancer incidence and mortality rates in US

AIGAL DIVYA

Index

Introduction:

The project is an extension of the Re-design project. In the re-design project, the focus was on to understand the effect of Sunlight, UV radiation and Melanoma Cancer incidences in the 51 states of the US. The main objective of this project would be to understand how are the factors like Race, Ethnicity and Gender related to the Melanoma incidences and mortality rates in the US. This relation is depicted through various visualizations and concepts learned in class over the duration of the semester.

1. Why this data-set?

The relation between Sunlight, UV radiation and Melanoma skin cancer incidences in the 51 US States was discussed in the re-design paper. From the Re-design, we deduced that the states of Texas, Florida California, New York have considerably higher rates of skin cancer incidences compared to other states. There could be various factors that affect the higher rates in some states. Race, ethnicity and gender could play an important role in the higher rates. This data-set was chosen in order to understand the relation between the above mentioned factors and Skin Cancer incidences/ mortality rates in the 51 states of US.

2. Data tables and Sources:

The focus of the data is to look for:

1) Incident count for various races and gender
2) Mortality count for various races and gender

The goal is to be able to compare the two factors, i.e. the incidences and the mortality counts. Hence, the data had to be gathered from different data-sets.

State	State Code	Age Group	Age Group C	Race	Race Code	Ethnicity	Ethnicity Coc	Sex	Sex Code	Count
Alabama	1	70-74 years	70-74	White	2106-3	Non-Hispanic	2186-5	Female	F	321
Alabama	1	70-74 years	70-74	White	2106-3	Non-Hispanic	2186-5	Male	M	704
Alabama	1	75-79 years	75-79	White	2106-3	Non-Hispanic	2186-5	Female	F	276
Alabama	1	75-79 years	75-79	White	2106-3	Non-Hispanic	2186-5	Male	M	618
Alabama	1	80-84 years	80-84	White	2106-3	Non-Hispanic	2186-5	Female	F	244
Alabama	1	80-84 years	80-84	White	2106-3	Non-Hispanic	2186-5	Male	M	470
Alabama	1	85+ years	85+	White	2106-3	Non-Hispanic	2186-5	Female	F	234
Alabama	1	85+ years	85+	White	2106-3	Non-Hispanic	2186-5	Male	M	327
Alaska	2	25-29 years	25-29	White	2106-3	Non-Hispanic	2186-5	Female	F	18
Alaska	2	35-39 years	35-39	White	2106-3	Non-Hispanic	2186-5	Female	F	22
Alaska	2	40-44 years	40-44	White	2106-3	Non-Hispanic	2186-5	Female	F	27
Alaska	2	40-44 years	40-44	White	2106-3	Non-Hispanic	2186-5	Male	M	22
Alaska	2	45-49 years	45-49	White	2106-3	Non-Hispanic	2186-5	Female	F	38
Alaska	2	45-49 years	45-49	White	2106-3	Non-Hispanic	2186-5	Male	M	56
Alaska	2	50-54 years	50-54	White	2106-3	Non-Hispanic	2186-5	Female	F	43

Data source 1: http://wonder.cdc.gov/cancer-v2012.HTML - Cancer incidence data

State	State Code	Age Group	Age Gr	Race	Race Cod	Ethnicity	Ethnicity Co	Sex	Sex Cod	Deaths
Alabama	1	70-74 year	70-74	White	2106-3	Non-Hispanic	2186-5	Female	F	78
Alabama	1	70-74 year	70-74	White	2106-3	Non-Hispanic	2186-5	Male	M	120
Alabama	1	75-79 year	75-79	White	2106-3	Non-Hispanic	2186-5	Female	F	52
Alabama	1	75-79 year	75-79	White	2106-3	Non-Hispanic	2186-5	Male	M	123
Alabama	1	80-84 year	80-84	White	2106-3	Non-Hispanic	2186-5	Female	F	52
Alabama	1	80-84 year	80-84	White	2106-3	Non-Hispanic	2186-5	Male	M	106
Alabama	1	85+ years	85+	White	2106-3	Non-Hispanic	2186-5	Female	F	76
Alabama	1	85+ years	85+	White	2106-3	Non-Hispanic	2186-5	Male	M	90
Alaska	2	60-64 year	60-64	White	2106-3	Non-Hispanic	2186-5	Male	M	17
Arizona	4	30-34 year	30-34	White	2106-3	Non-Hispanic	2186-5	Male	M	23
Arizona	4	35-39 year	35-39	White	2106-3	Non-Hispanic	2186-5	Male	M	23
Arizona	4	40-44 year	40-44	White	2106-3	Non-Hispanic	2186-5	Female	F	19
Arizona	4	40-44 year	40-44	White	2106-3	Non-Hispanic	2186-5	Male	M	40
Arizona	4	45-49 year	45-49	White	2106-3	Non-Hispanic	2186-5	Female	F	42

Data-source 2: http://wonder.cdc.gov/cancermort-v2012.HTML - Mortality data

The first data table gives us data for the incidence rates, while the second data gives us data for mortality rates with respect to the other factors of Gender, Race and Ethnicity.

3. Understanding the data:

As seen in the screenshot for the data tables, we understand that the table for Cancer incidence data contains various columns such as the name of the state, the state code, the columns for Ethnicity, Race, Age, Gender and the number of incidences. Details for various columns are listed as below:

a) ***Race*** - There are a total of 5 races. American Indian or Alaska Native, Asian or Pacific Islander, Black or African American, White, Other Races Combined and Unknown. The "Other Races Combined and Unknown" racial category contains data for the "Other" and "Unknown" racial categories.

b) ***Ethnicity*** – This data categorizes Ethnicity in two sections. 1) Hispanic and 2) Non-Hispanic

c) ***Age*** – Age is divided into the following categories:

Code	Description
1	< 1 year
1-4	1-4 years
5-9	5-9 years
10-14	10-14 years
15-19	15-19 years
20-24	20-24 years
25-29	25-29 years
30-34	30-34 years
35-39	35-39 years
40-44	40-44 years
45-49	45-49 years
50-54	50-54 years
55-59	55-59 years
60-64	60-64 years
65-69	65-69 years
70-74	70-74 years
75-79	75-79 years
80-84	80-84 years
85+	85+ years

d) ***Gender*** - Gender is categorized as 1) Male and 2) Female

4. Pre-processing data:

The previous releases of the data (1999-2005) included the "American Indian or Alaska Native" racial category within the "Other Races Combined" group. Due to this constraint, some states prior to 2005 did not have data for American Indian or Alaska Native separately. While after 2005, we can find the data properly segregated in the appropriate columns.

Also, there seemed to be an issue while fetching data from the Mortality data-set. When using all the filters for Race, Ethnicity, Gender and Age, the result seemed to ignore values for some categories for some states. But when using query for only 1 column, all relevant data was retrieved. Hence, a tedious route to gather data from each individual query and then merge it had to be taken.

During this process, a lot of data was gathered. To process all this data and to gain insights from it is a huge task, so the area of focus had to be narrowed and made more precise. In this process, the Age column was not considered for visualizations as there were a lot more categories for Age and deriving patterns considering all those categories would increase the complexity of the project manifold.

After ignoring data for the Age attribute, we limit our area of interest to Race, Ethnicity and Gender. In order to make the data more consistent and be able to analyze it efficiently, the data for Ethnicity and Race had to be categorized as one single column still maintaining the variety of the races so as to ensure that analysis of the result is not impacted. We name this column as Race-Ethnicity and categorize it into 4 levels. The data from Asian, Black and White remained the same, but data for the other 2 race categories under the Race column need to be aggregated. This new integrated column is named as “Hispanic” (Categories mentioned in the Design for Two factors example was referred for narrowing down the levels). Thus the 4 new Race-Ethnicity categories are as mentioned below:

- Asian
- Black
- Hispanic (Other Races Combined and American Indian included in this category)
- White

5. Data after pre-processing:

		Race and Ethnicity				Gender	
State	FIPS	Hispanic	Black	White	Asian	Female	Male
Alabama	1	0	0	9606	0	3916	5690
Alaska	2	0	0	621	0	264	357
Arizona	4	619	0	10572	0	4284	6907
Arkansas	5	0	0	4511	0	1828	2683
California	6	5049	52	70120	562	30870	44913
Colorado	8	0	0	10601	0	4600	6001
Connecticut	9	114	0	9408	0	4126	5396
Delaware	10	0	0	2349	0	941	1408
District of Co	11	0	0	194	0	36	158
Florida	12	1086	92	43642	0	17351	27469
Georgia	13	169	38	19332	0	8225	11314
Hawaii	15	0	0	2586	112	953	1745
Idaho	16	0	0	3686	0	1483	2203

6. Variables to be addressed:

The main focus of this project would be to answer the following questions:

1. What is the effect of **Race and Ethnicity** on cancer counts?

2. How does **Gender** relate to the cancer counts?
3. What is the proportion of Incidences and Mortality?

To find an answer for these questions, we shall require these 4 variables:

- Race-Ethnicity,
- Gender,
- Incidence and
- Mortality rates

7. Graphics and Visualizations used to analyze data:

a) Micromaps for Melanoma Cancer Incidence rate:

The re-design project focused on Micromaps, so for the final project the theme for the re-design project was continued and Micromaps are considered initially. This time a sorting is done on the basis of Melanoma cancer counts to understand the pattern of the states (western or eastern or nearer to the coast) have highest Melanoma cancer counts. ***(Figure 1: Micromaps sorted on Melanoma Cancer Incidences)*** below shows the data table presented in the form of a Micromaps for better visualization. From the Micromaps we understand that the states of Pennsylvania, Texas, New York, Florida and California show higher trend for skin cancer incidences as compared with other states.

b) *Dot-plot for comparison between Incidences and Mortality for various Races:*

Now that data for states having higher Incidence counts is in place, the focus is to find the relation of Incidences and Mortality counts for various races in the US. As mentioned earlier, the 4 different races are Asian, Black, Hispanic and White. From the plot, the Incidence and hence Mortality rates for White population are the highest (***Figure 2: Dot-plot for 4 races (Incidence & Mortality rates)***).

The Incidence and Mortality rates for the White population is so large that the rates for other races get subsided in the dot-plot. A second dot-plot is generated in which data for White population is removed and the focus is only on the other 3 races (Asian, Black and Hispanic). The second dot plot ***(Figure 3: Dot-plot for Asian, Black and Hispanic races***) shows that Although Hispanics have the highest Incidence rates, the Mortality rates for Black population is the highest for the 3 races.

c) Regression for White population (Gender-wise):

From the 2 dot-plots it is understood that, the Incidence and Mortality rates for White population is the highest. Hence, moving forward with the results obtained from the dot-plots, the focus now narrows down to only the White population. To understand the dynamics of Gender with respect to White population, regression is used to plot the Mortality counts for the Whites. One important thing to consider is that states-based incidence count relation was the focus of the re-design project. Hence, here going forward the area of interest would be only the data related to mortality rates.

Using regression, the data for White Females (X-axis) and White Males (Y-axis) is plotted ***(Figure 4:Mortality rates for Whites (Gender-wise) using Regression***). The regression plot uses linear model. In the regression plot, the blue line is the fitted line. The confidence interval, the model co-efficient, predictions from result of the model and residual values for linear model are calculated for the data. A ggplot is then used to plot the values obtained for the linear model. As can be seen from the regression analysis, the regression line has a positive trend. Also, it can be noticed, that scale values for Male starts from 0 extending all the way to 6000, while the scale values for Females extends to 3000 starting from 0. The maximum for males is seen to be a value more than 6000, while the maximum for Females shows to be around 3000. From the regression analysis, we can see that the mortality counts for White males is almost double to that seen for White females.

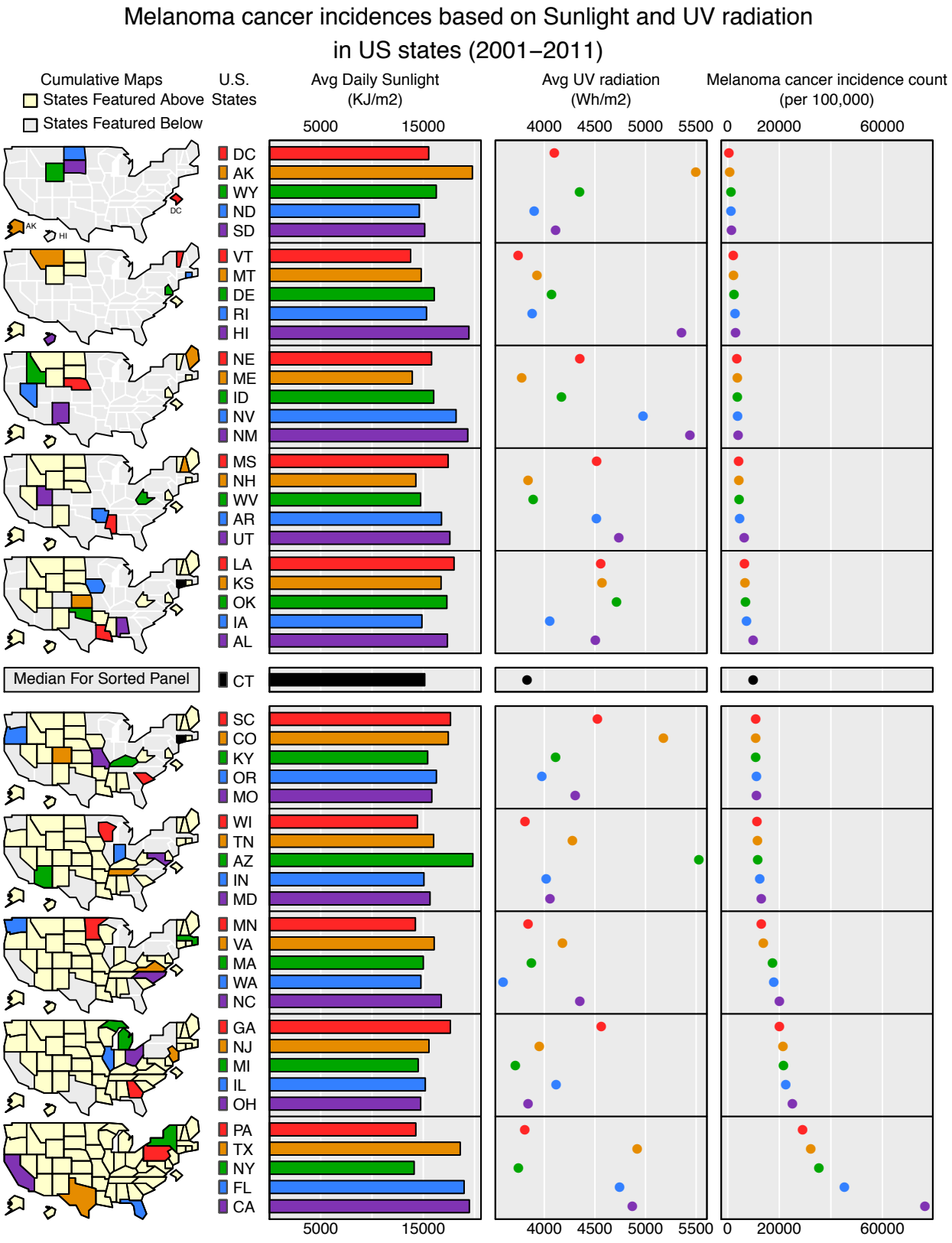

Figure 1: Micromaps sorted on Melanoma Cancer Incidences

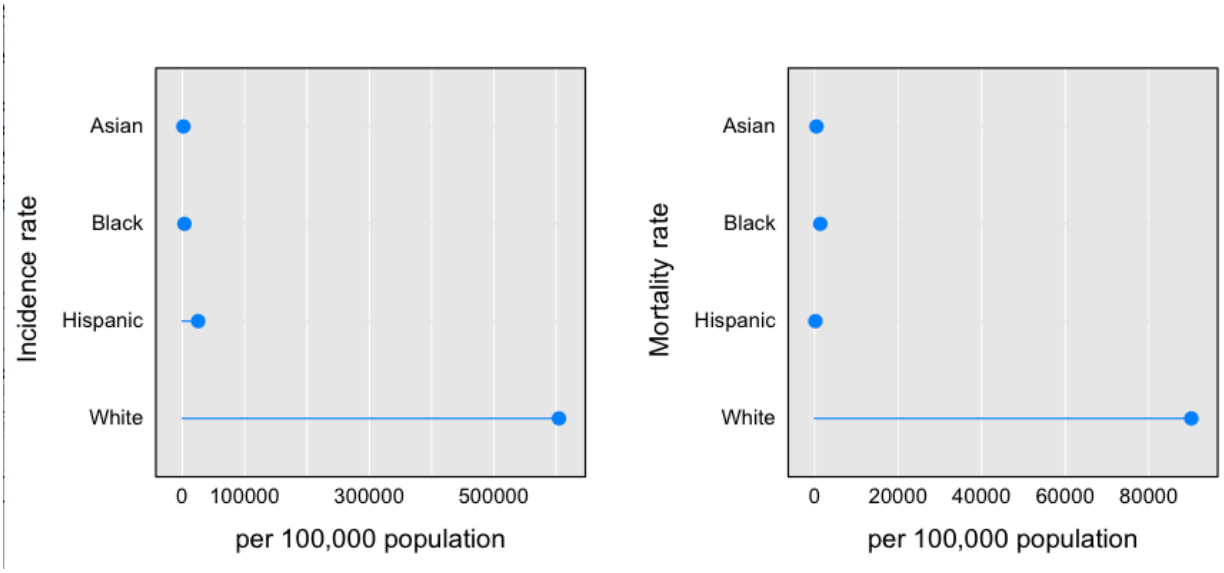

Figure 2: Dot-plot for 4 races (Incidence & Mortality rates)

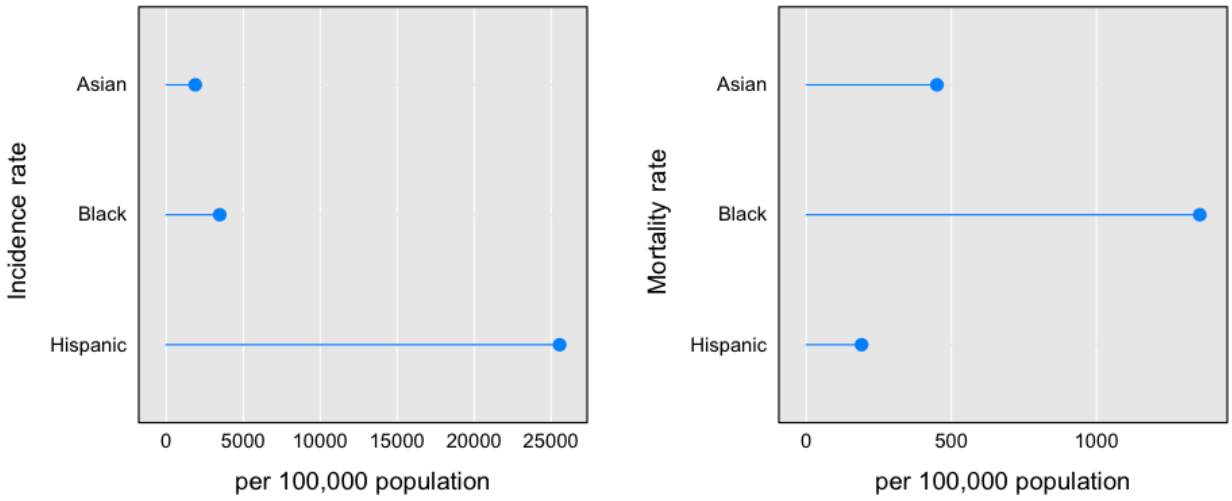

Figure 3: Dot-plot for Asian, Black and Hispanic races

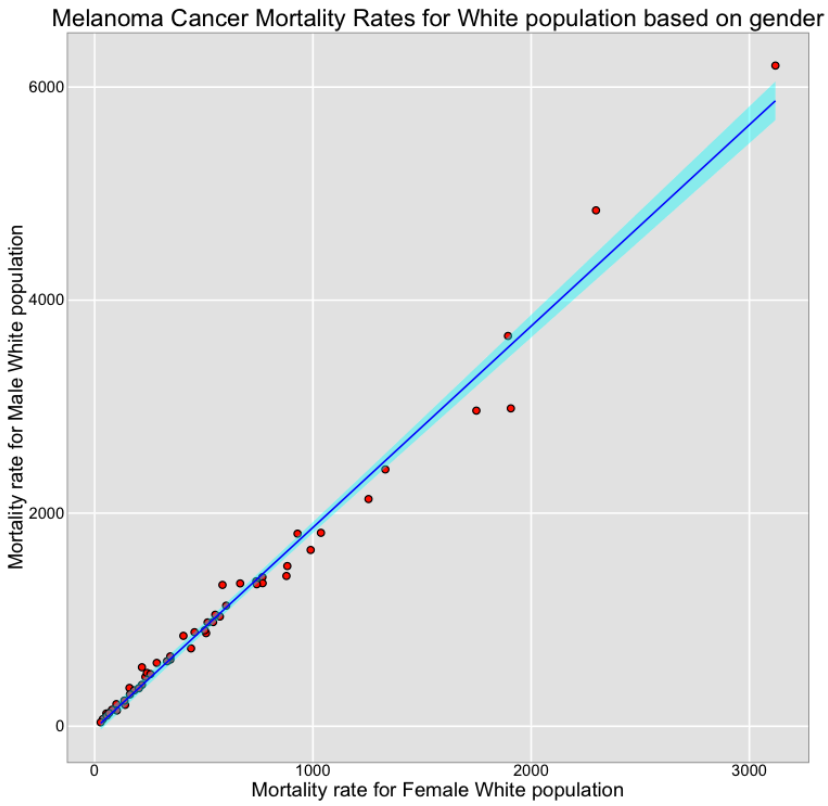

Figure 4:Mortality rates for Whites (Gender-wise) using Regression

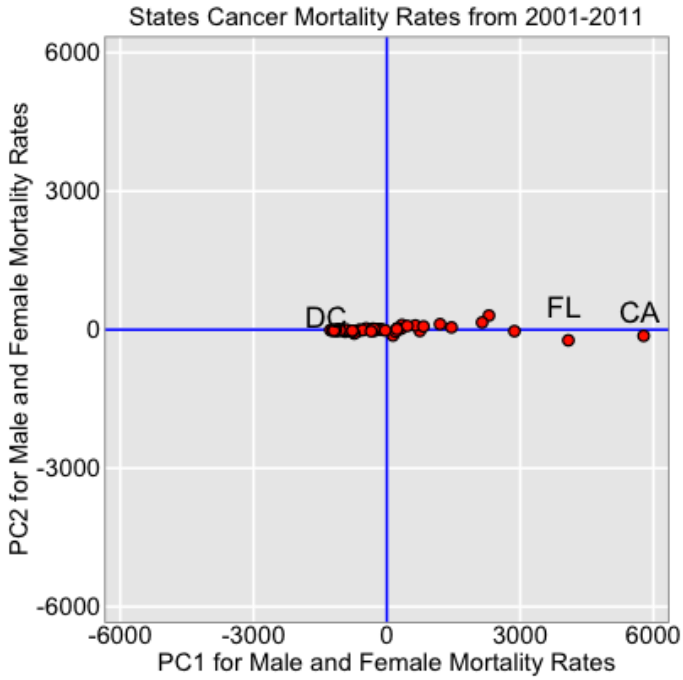

Figure 5: Principal Component Analysis for Mortality counts based on Gender data

d) Principal Component Analysis for Mortality rates (Gender data):

Mortality counts for both genders need to be understood for the US states. The data available is from 2001 to 2011, so it could be interesting to understand the trend for mortality counts based on Gender for this duration. Principal Component Analysis is used to understand these dynamics.

The first step in constructing principal components is to center the variables (subtract each variable's mean so the new mean is zero). Then scaling is done if units used are different(prcomp() function is used). Our data was in same units. Scaling divides the centered variables by their standard deviations which forms the Correlation matrix. The next step selects the principal components with largest variances for further analysis. The smallest number of components that combined have at least 80% of the variance relative to the sum of variances for all the original variables is used. PC1 gives the sum of variances for Male and Female. The values for PC1 is 99% and that for PC2 is 100%.

The PCA plot is then plotted using ggplot (***Figure 5: Principal Component Analysis for Mortality counts based on Gender data***). PC1 values for male and female mortality rates are displayed on the X-axis, while PC2 values are presented on the Y-axis. The plot shows vividly a nice cluster of red dots. This cluster (the red dots) depicts those states which have nearly similar values. As can be seen from the plot, some red dots stand out from the cluster. For understanding better, some data points are labelled with their state abbreviations. 3 extreme value data points are shown on the plot. California has maximum while DC has minimum value for PC1. Also, Florida has the minimum value for PC2.

e) Biplot for PC1 and PC2 for Gender data:

Moving ahead, a biplot is used to represent gender data more aptly. The PCA using ggplot manifested data in the form of a cluster of red dots, which symbolized the states in the US. A biplot would similarly show the PCA values (PC1 and PC2) for gender data with the help of state abbreviations.

The biplot uses points to represent the scores of the states on the principal components, and it uses vectors to represent the coefficients of the variables on the principal components. It can be considered as an arrow plot from the origin with the top and right axes providing the coordinate system. The axes for the biplot are a pair of principal components. We plot PC1 on X-axis and PC2 on the Y-axis (***Figure 6: Biplot showing PC1 and PC2 for Gender data***).

After plotting the PC values, the plot shows a nice cluster near about at the center. As could be understood from the biplot, points in the cluster that are close together correspond to observations that have similar scores. Points that stand out from the cluster depict those states having mortality rate that vary largely from the rates for the cluster. Again, we can notice that the states of Pennsylvania, Texas, New York, Florida and California deviate far from the rest of the states.

f) Ordering states based on PC values:

For Principal Component Analysis, the values for PC1 and PC2 are computed. All 51 states are then ordered based on these values to understand the mortality dynamics in all the states more clearly (Figure 7: Ordering states based on PC1). The states with the 5 least and 5 highest values are highlighted in the table. The first 5 states (least PC values) as seen from the table are (DC, AK, ND, WY and VT) while the states with 5 highest PC values are (PA, NY, TX, FL, CA)

g) Micromaps for Melanoma Cancer Mortality based on Gender in the US states:

In order to check the correctness of the values computed by PCA, micromaps are used to visually verify the results. A micromaps for mortality rates for all the US states based on gender data is depicted in ***(Figure 8: Micromaps for Gender data***). The gender data is from 2001-2011. From the micromaps, the 5 states with least cancer mortality rates are DC, AK, ND, WY and VT and the highest rates are found in the states of PA, NY, TX, FL, CA. This verifies the analysis done by PCA. The ordering of the states on the basis of PCA and micromaps is analogous.

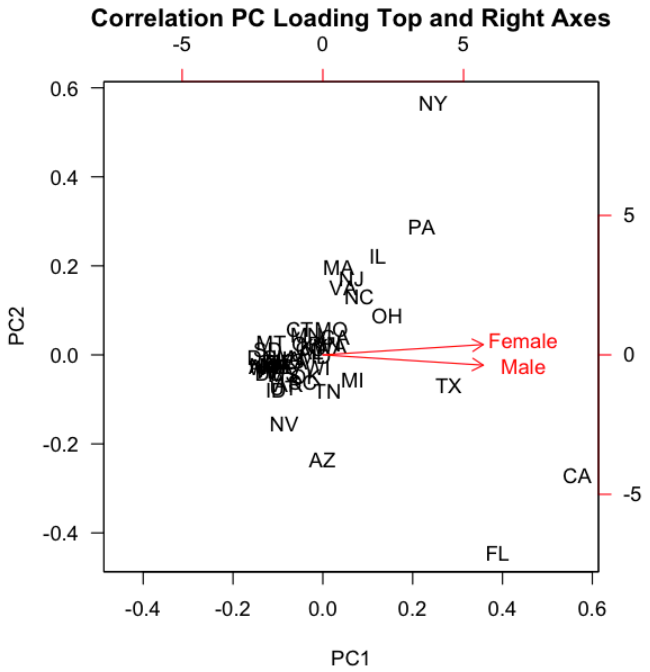

Figure 6: Biplot showing PC1 and PC2 for Gender data

Sr.No	States	PC1	PC2
1	DC	-1270.3	-6.7
2	AK	-1246.4	-17.2
3	ND	-1215.9	-15.2
4	WY	-1191.4	-18
5	VT	-1156.5	-20.7
6	SD	-1151.5	3.8
7	HI	-1142.7	-22.4
8	DE	-1098.9	-25.2
9	MT	-1082	11
10	RI	-1053	-11.1
11	NH	-990.9	-15.8
12	ME	-946.2	-16.9
13	ID	-937.5	-45.3
14	NM	-913.3	-6
15	NE	-883	-10.8
16	MS	-783.1	-25.3
17	WV	-777.2	-22.8
18	UT	-775	-42.1
19	NV	-733.1	-84.7
20	AR	-660.4	-37.2
21	KS	-632	-9.3
22	IA	-612.7	-7.1
23	LA	-544.3	-10.5
24	CT	-466.9	28.7
25	SC	-353.5	-34.9
26	MN	-311.8	22.3

Sr.No	States	PC1	PC2
27	OK	-294.2	-27.8
28	OR	-288.8	12.3
29	CO	-202.6	6.2
30	AL	-195.5	2.4
31	KY	-140.8	9.6
32	MD	-105.3	-2.5
33	WI	-35.4	-15.8
34	AZ	-137.7	126.9
35	TN	195.2	-44.2
36	WA	225.2	10.1
37	MO	238.2	30.3
38	IN	238.2	13.3
39	GA	320.3	20.7
40	MA	342	104.3
41	VA	460.8	80.8
42	NJ	640.9	91.7
43	MI	742.2	-28.9
44	NC	824.3	70.2
45	IL	1196.8	119.3
46	OH	1455	49
47	PA	2143.8	156.6
48	NY	2296.9	306.3
49	TX	2871.1	-32.5
50	FL	4084.8	-230.7
51	CA	5778.6	-134.6

Figure 7: Ordering states based on PC1

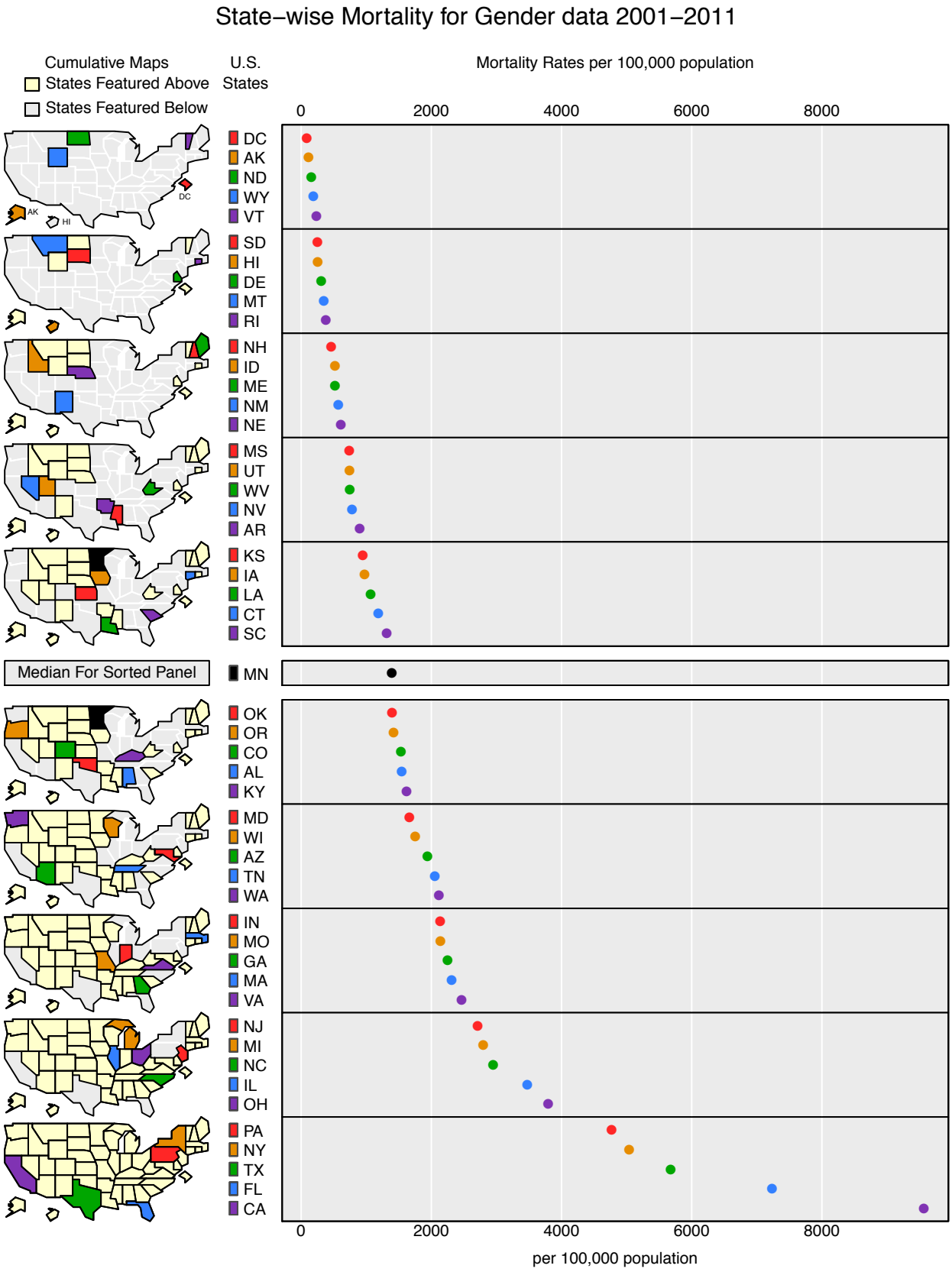

Figure 8: Micromaps for Gender data

8. Results and Conclusions:

The focus of the final project was to understand the dynamics of attributes like Race, Ethnicity and Gender on melanoma cancer incidence and mortality rates. This was done with the help of various visualizations starting from Micromaps, Dot-plots and then moving towards complex models like Regression, Principal Component Analysis and Biplots. There was also an attempt to check the similarity of analysis done by Principal Component Analysis and that from Micromaps. This is handled towards the end of the paper.

We could infer the following from all the visualizations and analysis done so far:

1. Incidences and hence Mortality rates are higher among White population.
2. The states of CA, NY, TX, FL, PA have higher counts for Melanoma mortality compared to the rest of the states in the US.
3. White population from the states of CA, NY, TX, FL, PA have greater chances of mortality than rest of the population.
4. White Males have a higher probability to get Melanoma cancer than White Females.

Some interesting statistics to note:

While working on this project, the following interesting statistics were revealed:

1. 1 in 50 men and women diagnosed with melanoma of the skin during their lifetime
2. Melanoma accounts for less than two percent of skin cancer cases, but the vast majority of skin cancer deaths.
3. Average survival with melanoma increased from 49 percent (1950 – 1954) to 91 percent today
4. Regular daily use of an SPF 15 or higher sunscreen reduces risk of developing melanoma by 50 percent. [1]

9. References:

1. *http://www.skincancer.org/skin-cancer-information/skin-cancer-facts#ethnicity*
2. *Micromaps, Biplot, PCA, Regression – Professor Daniel Carr, George Mason University*

10. Acknowledgements:

Submitted as a Final Project for STAT 515 course under Dr. Daniel Carr, George Mason University