Lab 9 – Modeling Melanoma Rates

NAME 1 – NETID

NAME 2 – NETID [if applicable]

NAME 3 – NETID [if applicable]

Open the following short article about melanoma rates across U.S. States and read the first three paragraphs: https://www.cancercenter.com/community/blog/2016/05/unhealthiest-states-for-skin-cancer-risk-may-surprise-you

As you'll notice, this article is considering different possible factors that might explain why some states have higher melanoma rates than others. One particular factor is suggested by Dr. Charles Komen Brown, where he notes **that states with lower sun exposure on average tend to have higher melanoma rates.** He offers one possible *causal* explanation by saying *it is because residents in this state are not used to thinking about sun exposure and skin protection*. In this lab, we will investigate this relationship (among others) and also consider the possibility that this is truly a causal link. We're going to do some modeling for explanation!

The *Melanoma.xlsx* file contains state data collected through various sources (most of which is from 2014-2017). Note this data includes the 48 contiguous states (excludes Alaska and Hawaii). The variable descriptions are below

READ THIS

Name: State Name

Abb: State Abbreviation

Mel: Number of new melanoma cases per 10,000 people each year

Sun: The average sun exposure in each state, measured in kilojoules per square meter (kJ/m^2)

Lat: Latitude of the center point of the state

Lon: Longitude of the center point of the state

Ocean: Whether the state borders an ocean or not

Temp: Average daily high temperature year around

Pop: State Population

Density: Average number of state residents per square miles

Age: Average age of state residents

White: Proportion of state residents that identify as both "white" and "non-hispanic"

Income: Average income of state residents

Step 0

- Upload and import the Melanoma.xlsx file into RStudio
- Open up a new script (or RMarkdown file) to work in.
- **Do you remember what package you should library before getting started?**
- Note, we won't use many of these variables, but they are here if you want to explore!

(5pts) Question 1: Let's start by creating a map of the United States and coloring each state by the Melanoma Rate. This should give us a little more insight to where melanoma rates are high.

Start by installing and librarying the package `"maps"`.

```
install.packages("maps")
library(maps)
```

This package contains quite a bit of map data, but we're only going to be working on map data from U.S. states.

Before doing this next step, make sure you have libraried tidyverse as well! This next step pulls from data in maps, but uses a tidyverse function.

This next step will call up some U.S. state map border data—let's save it as a dataframe called "MainStates." Feel free to click on it in your global environment to take a look!

```
MainStates = map_data("state")
```

Next, we want to merge this data frame into our Melanoma data frame. But to do that, we need a common column. Our `"Name"` column in Melanoma contains state names, but in MainStates, it's called `"region"`. Run the following command to change the MainStates data frame column to `"Name"`. Then, we will be able to easily match up these columns when we merge the data frames.

```
names(MainStates)[names(MainStates) == "region"] = "Name"
```

Now we're ready to merge! Run the following function (from the dplyr package) to merge our data. Go ahead and save this merged data frame under a different name (suggested Melanoma_Map) because the dimensions are going to radically change, and it will be nice to still have the original available in this session!

```
        Melanoma_Map = inner_join(MainStates, Melanoma, by = "Name")
```

We're finally ready to create the map! Here are your instructions:

- Make a ggplot with Melanoma_Map as your data
- In the aes line, set x = long, y = lat, group = group, and fill = Mel
    - Note, these are variable names from *Melanoma_Map*, not from Melanoma
- Use geom_polygon(), and set color = "black" inside this argument to insert black state borders
- Use a fill color palette (with distiller) to choose a custom color palette. The default palette is blues, but let's take this opportunity to pick a palette with more dynamic differences on each side of the scale!
- Add theme_classic() to create a blank background
- Add an appropriate title

Insert your map graph

Which two states stand out as having the highest melanoma rates?

(4pts) Question 2: Next, let's look at the relationship between a state's average sun exposure and a state's melanoma rate. Create a scatterplot using sun exposure as the predictor variable and melanoma rate as the response variable. However, *in place* of using geom_point, substitute in the following geom code:

```
geom_text(aes(label=Abb),hjust=0, vjust=0, size = 3, fontface = 2)
```

In addition:

- Add a best fit line (Standard error shading optional, Color choice up to you)
- Add a title
- Adjust the x and y axis titles to be fully written in 1-3 words (rather than the default variable abbreviations)
- Use the theme_classic() theme style (to provide a blank background)

Please include an image of this scatterplot in the report (code optional)

Which states have higher melanoma rates—those with more sun exposure on average, or less sun exposure?

Which one or two states appear to have the largest residuals in this model?

(5pts) Question 3: Now, create a simple linear model with sun exposure as a predictor of melanoma rate.

Copy the <u>summary output</u> of this model into your report *(starting with the heading that says "Coefficient" through to the end).*

Interpret the multiple r-squared value *(use the fill in the blank template we learned in chapter 12)*

Is there evidence that average state sun exposure and state melanoma rate have at least some linear association? *(Don't just say "reject" or "fail to reject" the null hypothesis—try using the interpretational language we learned in Chapter 8. Little to no evidence, moderate evidence, strong evidence, very strong?).*

(4pts) Question 4: Let's look at some other possible predictors of melanoma rate: Average temperature, Average age, and Proportion of white, non-hispanic residents. Create a simple linear model for each of these numeric predictors (i.e., create three simple linear models) and look at the output.

Are any of these three variables stronger predictors of melanoma rate than sun exposure? Explain.

Create a scatterplot of the strongest predictor from these three. Again, use state abbreviations in place of points. Follow all formatting guidelines from question 2. **Include this plot in your report.**

(4pts) Question 5: While it is possible to create a model with two numeric predictors, let's focus on the simpler case of using one numeric predictor with one binary predictor.

Using an `ifelse` statement, create a new variable in the dataset called `White_Bin` that records "High" if the state's proportion of residents being "white, non-hispanic" is above 0.7 and "Low" if the proportion is below 0.7.

Create a scatterplot with Sun exposure on the x axis, color the data points based on whether the proportion of white residents is above 0.7 or not, and put Melanoma rate on the y axis.

- Continue to use the state abbreviations in place of points.
- Add best fit lines for each group (i.e., visualize what an interaction model would look like).
- Add an appropriate title
- Adjust the axes labels. Also adjust the color legend to say "Proportion White." You can do that by entering a color = "…" argument into the labs function.

Include your code *and* an image of your plot here

(4pts) Question 6: In the previous question, we plotted an interaction model (allowing for different slopes), but let's explore some different model options. First, create an additive model for Sun exposure and Proportion White (binary version) as predictors of Melanoma rate. Then create an interaction model with the same variables.

> **Include your additive model summary output and your interaction model summary output in your report**
>
> **Is there evidence that we gain anything by including an interaction term between sun exposure and proportion white, or is it more sensible to allow proportion white to contribute only as an additive term?** *Justify your answer using your model summaries.*

(4pts) Question 7: Consult your **additive** model from Question 6 *(looking at your plot from Question 5 might help too!)* and answer these questions:

> **After identifying whether a state has above or below 70% white, non-hispanic residents, is there evidence that sun exposure has a linear relationship with melanoma rate?** *Briefly justify*
>
> **If comparing two states with the same avg. sun exposure, is there evidence that proportion of white and non-hispanic residents is a predictor of melanoma rate?** *Briefly justify*
>
> **Contextually, how might the proportion of white residents act as a *confounder* to the association between sun exposure and melanoma rate?** *(Check back to Chapter 2 and look for the confounding diagram. Think about what that might look like here!).*