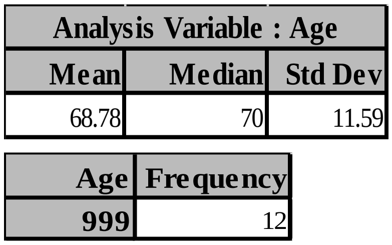
Caleb Fornshell

I am using a subset of the 1975-2016 RESPIR data file from the SEER datasets. I am only using the years 1988-2015, because that allowed me combine two different variables that both measured tumor size. The variables I am using are: Age at Diagnosis, Size of Tumor, Survival Months, Sex, Race, Year of Diagnosis, State(via registry information), Marital Status, Tumor Behavior, and Insurance Status. The first three variables are quantitative, so I will provide basic descriptive statistics and histograms of the variables. The last seven variables are categorical, so I will provide bar graphs and count and frequency data for the categories in in each variable. I included the variable name from the SAS read in file after the bolded header for each variable.

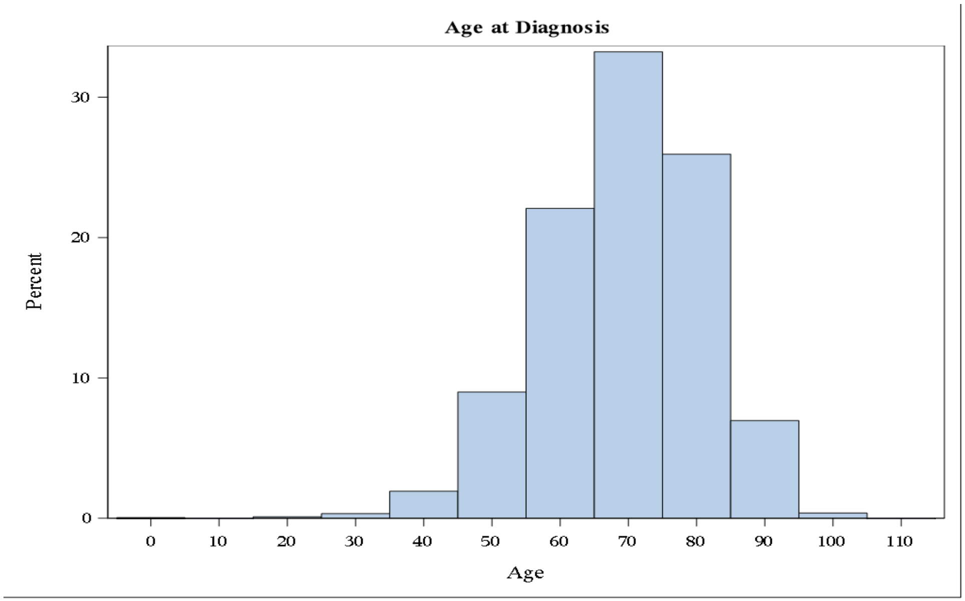
Age at Diagnosis: (age_dx)

The Age variable appears symmetric in shape. The results are not surprising. Relatively older people make up most of the cases. I removed observations with missing data indicated by a value of 999.



Analysis Variable : Age		
Mean	Median	Std Dev
68.78	70	11.59

Age	Frequency
999	12

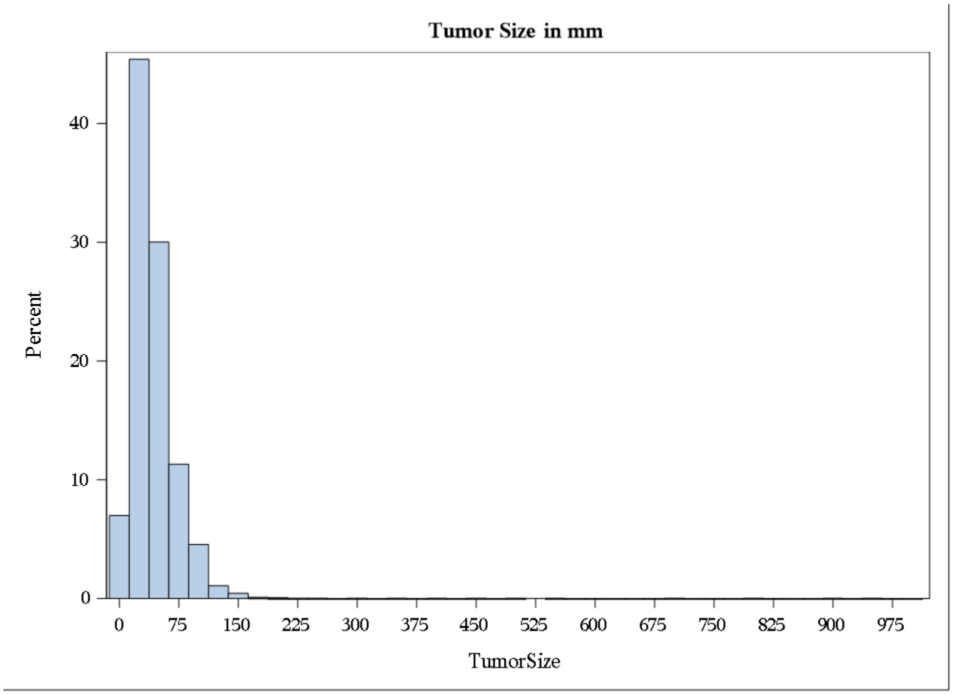


Tumor Size: (eod10_sz and cstumsize)

This variable was a combination of two different variables in the SEER data: eod10_sz and cstumsize. Both variables measured tumor size, but they consisted of different years. Eod10_sz consisted of the years 1988-2003, and cstumsize was for the years 2004-2015. I had to remove observations of measurements greater than 988, because the observations had values that were not precise, or the size was unknown. For example, measurements of 991-995 had measurements of the form “less than X mm”, and 999 indicated a missing value. The omitted values are shown in the second table. All measurements are in millimeters. Most measurements were less than 100 mm, but there were some larger values. This caused the data have right skew.

Analysis Variable : TumorSize		
Mean	Median	Std Dev
41.81	35	30.59

TumorSize	Frequency
989	29
990	217
991	93
992	175
993	245
994	134
995	65
996	130
997	116
998	512
999	184505

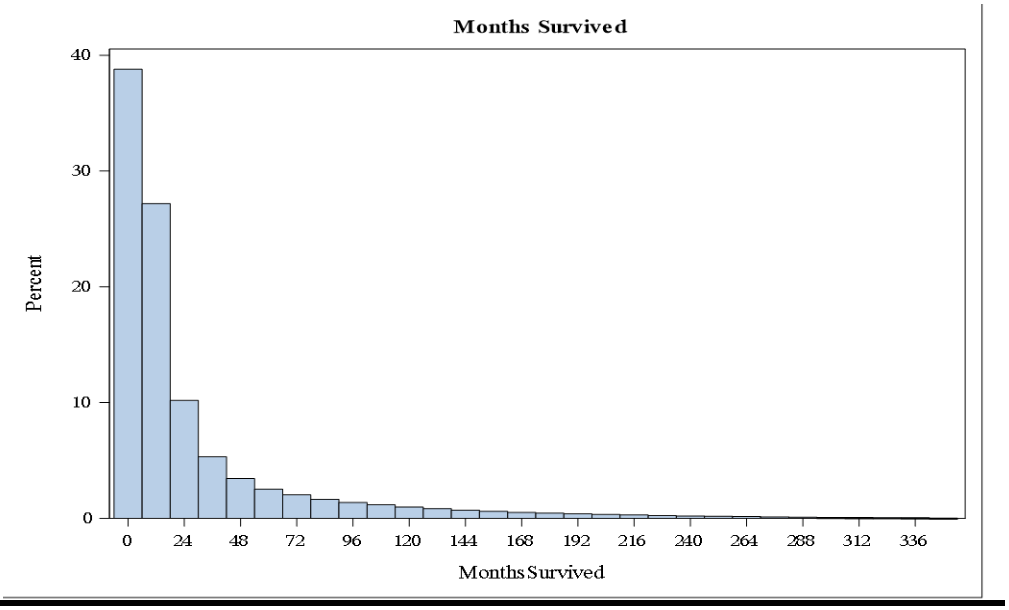


Months Survived: (srv_time_mon)

I removed observations that were coded with a 9999 as these values indicated the value was missing. Most individuals had measurements of a few months to few years, but there were many that survived longer. This variable is also right skewed.

Analysis Variable : MonthsSurvived		
Mean	Median	Std Dev
27.48	9	47.21

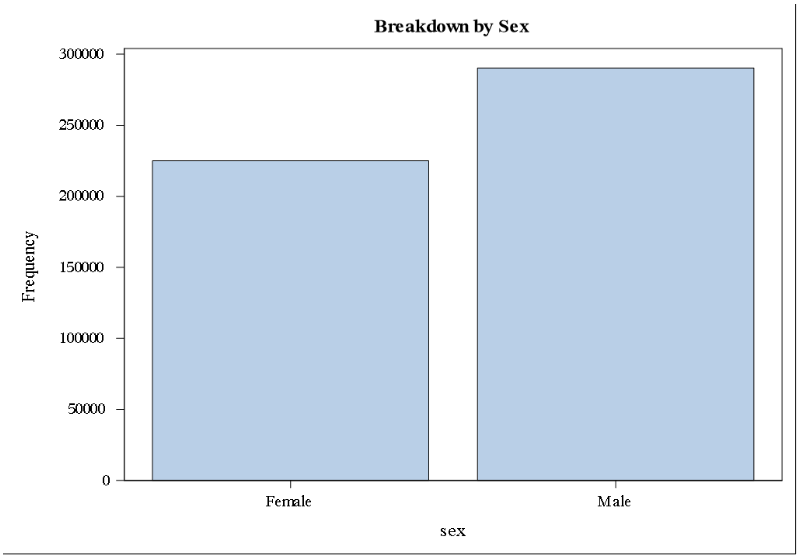
MonthsSurvived	Frequency
9999	11258



Sex: (sex)

The majority of cases were from males.

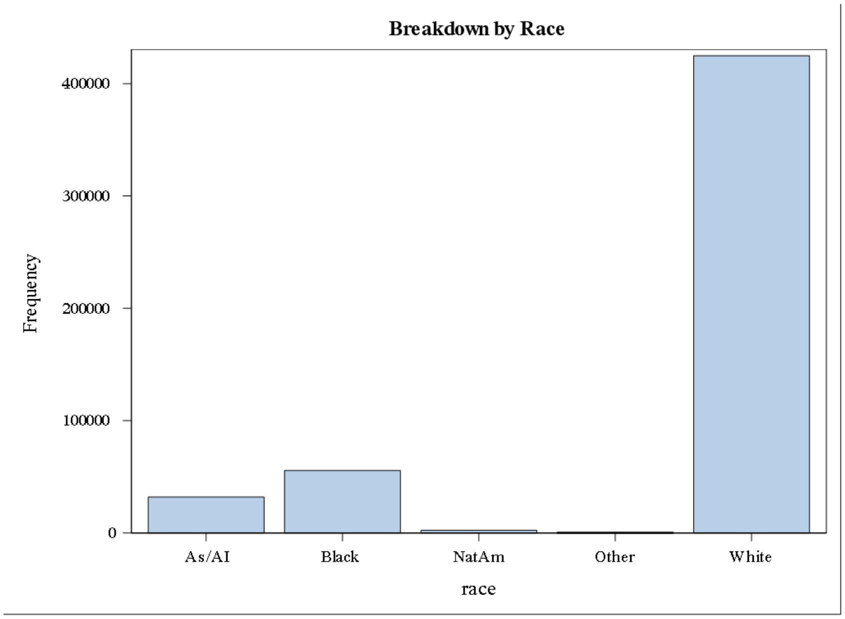
sex	Frequency	Percent
Female	224975	43.66
Male	290339	56.34



Race: (rac_recy)

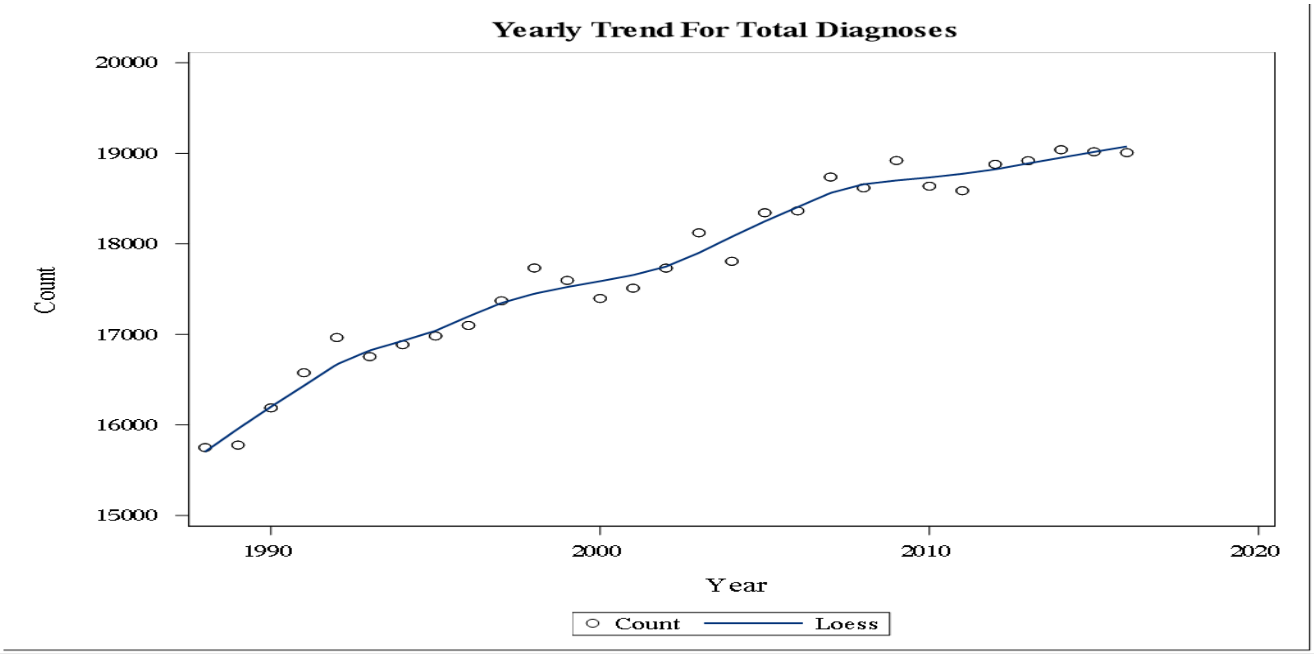
The most common race represented was by far white. The information came from the Race Recode, “Rac_Recy” variable. I combined the categories for “other” and “unknown”. As/AI stands for Asian and Asian Islander.

race	Frequency	Percent
As/AI	31968	6.2
Black	55504	10.77
NatAm	2247	0.44
Other	719	0.14
White	424876	82.45



Year: (created variable of all 1s and grouped by year_dx)

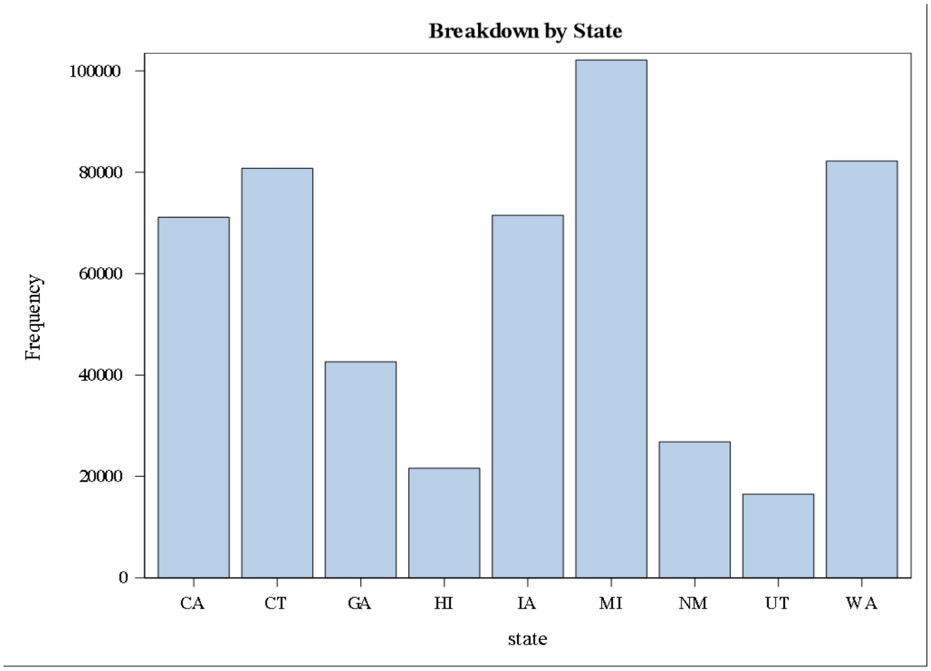
There was an overall upward trend. This could be explained by increasing population or more sites reporting.



State: (reg, recoded to state)

There does not appear to be any pattern among the states. This information was based on the Registry ID variable.

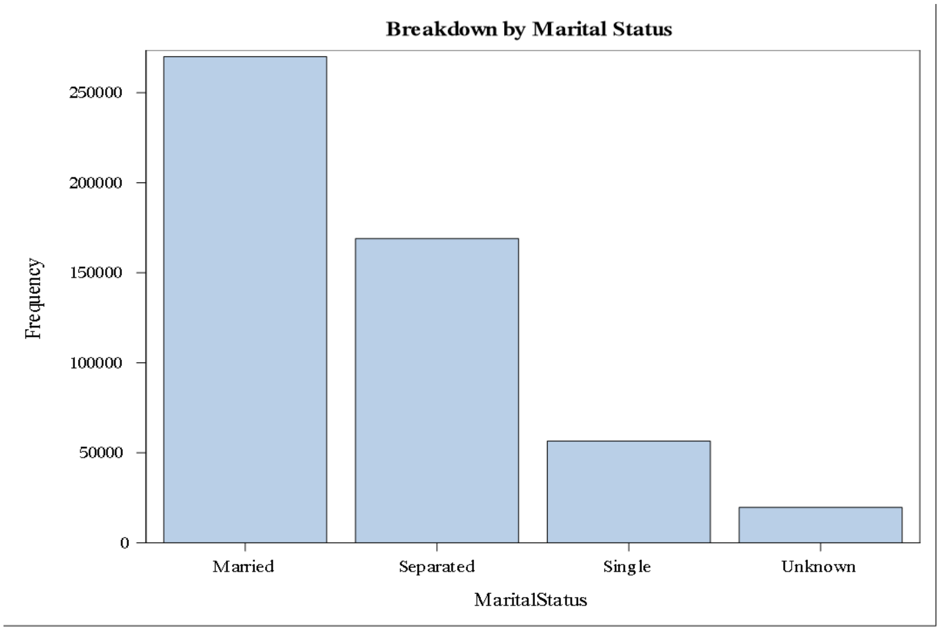
state	Frequency	Percent
CA	71111	13.8
CT	80765	15.67
GA	42612	8.27
HI	21621	4.2
IA	71526	13.88
MI	102130	19.82
NM	26812	5.2
UT	16508	3.2
WA	82229	15.96



Marital Status: (mar_stat)

Most of the diagnosed were married at the time of their diagnosis. I combined the categories that indicated the individual was once married but no longer is. For example, divorced, widowed, and separated are all in the "Separated" category. Single means the individual was never married.

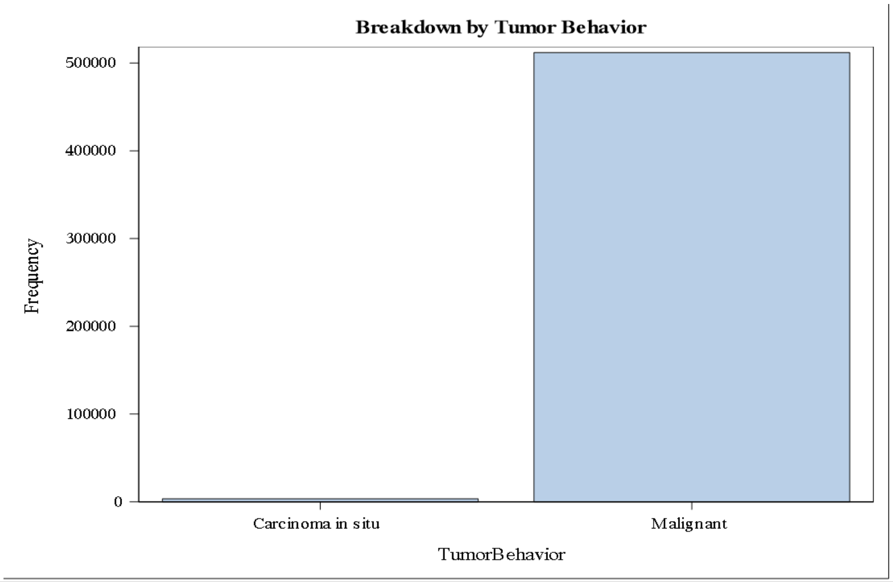
MaritalStatus	Frequency	Percent
Married	270090	52.41
Separated	169035	32.8
Single	56545	10.97
Unknown	19644	3.81



Tumor Behavior: (beho3v)

This variable was dominated by the malignant category. This variable is supposed to have four different categories, but only two are present in this dataset.

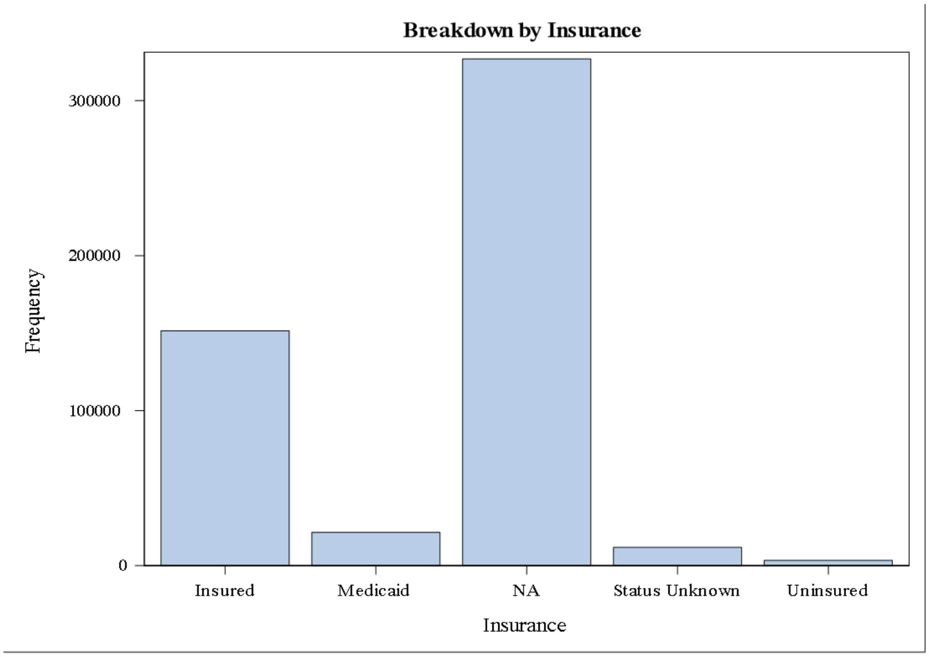
TumorBehavior	Frequency	Percent
Carcinoma in situ	3429	0.67
Malignant	511885	99.33



Insurance Status: (insrec_pub)

Most people who were diagnosed had some sort of insurance. The large number of NA values is because this variable only counts the observations from 2007 or later.

Insurance	Frequency	Percent
Insured	151587	29.42
Medicaid	21521	4.18
NA	326953	63.45
Status Unknown	11808	2.29
Uninsured	3445	0.67



SAS Code:

```
/*SAS code for STAT 770 Homework 1. Chose 10 variables from SEER data and give brief
description of each using descriptive statistics and frequencies, counts ect.
Using subset of RESPIR SEER data. I left out the SAS read in file included with the data files.
Due 1/27/2020 */


/*Selecting Variables and coercing to proper data type*/
proc sql;
        CREATE TABLE RESPIR_VARS as
        SELECT pubcsnum as ID,
                input(reg, 10.) as Registry,
                sex as Sex_,
                mar_stat as MaritalStatus_,
                input(age_dx, 4.) as Age,
                input(year_dx, 4.) as Year,
                input(eod10_sz, 4.) as TS88_03,
                input(cstumsiz,4.) as TS04_15,
                rac_recy as Race_,
                input(srv_time_mon, 4.) as MonthsSurvived,
                beho3v as TumorBehavior_,
                insrec_pub as Insurance_
        FROM RESPIR;
QUIT;

/*Combining Tumor size variables and subsetting data based on year
Making other variables easier to interpret(Sex, registry, race) */
data RESPIR_VARS;
        set RESPIR_VARS;
        one=1;
        TumorSize=0;
        if TS88_03>TS04_15 then TumorSize = TS88_03;
        else TumorSize=TS04_15;
        where Year>=1988 ;
        drop TS88_03 TS04_15;

        /*       Coding race */
        race="-----";
        if race_ = 1 then race="White";
        else if race_ = 2 then race="Black";
        else if race_ = 3 then race="NatAm";
        else if race_ = 4 then race="As/AI";
        else race="Other";

        /*Coding State Based on Registry Id */
        state="--";
        if  registry in (1501, 153, 1535, 1541) then state="CA";
```

```
	else if  registry = 1502 then state = "CT";
	else if  registry = 1520 then state = "MI";
	else if  registry = 1521 then state = "HI";
	else if  registry = 1522 then state = "IA";
	else if  registry = 1523 then state = "NM";
	else if  registry =  1525 then state = "WA";
	else if  registry =  1526 then state = "UT";
	else if  registry in (1527,  1537,  1547) then state = "GA";
	else if  registry =  1529 then state = "AK";
	else if  registry =  1542 then state = "KT";
	else if  registry =  1543 then state = "LA";
	else if  registry =  1544 then state = "NJ";
	else state="other";

	/*Coding Sex*/
	sex="------";
	if sex_="1" then sex = "Male";
	else sex="Female";

	/*Coding Tumor Behavior*/
	TumorBehavior="-------------------";
	if TumorBehavior_="0" then TumorBehavior="Benign";
	else if TumorBehavior_="1" then TumorBehavior="Mal Pot.";
	else if TumorBehavior_="2" then TumorBehavior="Carcinoma in situ";
	else if TumorBehavior_="3" then TumorBehavior="Malignant";
	else TumorBehavior="NA";

	/*Coding insurance*/
	Insurance="--------------";
	if Insurance_="1" then Insurance="Uninsured";
	else if Insurance_="2" then Insurance="Medicaid";
	else if Insurance_ in ("3", "4") then Insurance="Insured";
	else if Insurance_="5" then Insurance="Status Unknown";
	else Insurance="NA";

	/*Recoding Mariage Status*/
	MaritalStatus="---------";
	if MaritalStatus_="1" then MaritalStatus="Single";
	else if MaritalStatus_="2" then MaritalStatus="Married";
	else if MaritalStatus_ in ("3","4","5","6") then MaritalStatus="Separated";
	else MaritalStatus="Unknown";

	drop race_ Insurance_ Sex_ TumorBehavior_  MaritalStatus_;
run;
```

```
**************************************************;
/* Descriptive Statistics for Age at Diagnosis  */
**************************************************;
title2 "Age at Diagnosis";
proc means data=RESPIR_VARS mean median std maxdec=2;
	var age;
	where age<999;
run;

/* Showing number of observations that were removed due to missing obs */
proc freq data=RESPIR_VARS;
	table age /  nopercent nocum;
	where age>=120;
run;

/* Histogram for age*/
proc sgplot data=RESPIR_VARS ;
	histogram age / binstart=0 binwidth=10 showbins;
	where age<999;
run;


**************************************************;
/* Descriptive Statistics for Tumor Size         */
**************************************************;
title2 "Tumor Size in mm";
proc means data=RESPIR_VARS mean median std maxdec=2;
	var TumorSize;
	where TumorSize<989;
run;

/* Showing number of observations that were removed due to imprecise measures  */
proc freq data=RESPIR_VARS;
	table TumorSize /  nopercent nocum;
	where TumorSize>=989;
run;

/* Histogram for Tumor size*/
proc sgplot data=RESPIR_VARS ;
	histogram TumorSize / binstart=0 binwidth=25 showbins;
	where TumorSize<989;
run;
```

```
*****************************************************;
/* Descriptive Statistics for Survival Months    */
*****************************************************;
title2 "Months Survived";
proc means data=RESPIR_VARS mean median std maxdec=2;
	var MonthsSurvived;
	where MonthsSurvived<9999;
run;

/* Showing number of observations that were removed due to imprecise measures  */
proc freq data=RESPIR_VARS;
	table MonthsSurvived /  nopercent nocum;
	where MonthsSurvived>=9999;
run;

/* Histogram for Survival Months*/
proc sgplot data=RESPIR_VARS ;
	histogram MonthsSurvived / binstart=0 binwidth=12 showbins;
	where MonthsSurvived<9999;
run;

*************************************************
/* Counts and relative frequencies for Sex */
*************************************************;
title "Breakdown by Sex";
proc freq data=RESPIR_VARS;
	table Sex / nocum;
run;

/* barplot for Sex */
proc sgplot data=RESPIR_VARS;
	vbar Sex;
run;


**************************************************
/* Counts and relative frequencies for Race */
**************************************************;
title "Breakdown by Race";
proc freq data=RESPIR_VARS;
	table Race / nocum;
run;

/* barplot for Race */
proc sgplot data=RESPIR_VARS;
	vbar Race;
run;
```

```
******************************************
/* Counting number of obs for each year */
******************************************;
Title "Yearly Trend For Total Diagnoses";
proc sql;
	create table years as
	select year, sum(one) as Count
	from RESPIR_VARS
	group by year;
quit;

/* Using counts of diagnosis by year for trend of diagnosis every year */
proc sgplot data=years;
	scatter x=year y=count;
	loess x=year y=count;
	xaxis  min=1988 max=2018;
	yaxis  min=15000 max=20000;
run;


***********************************************
/* Counts and relative frequencies for State */
***********************************************;
title "Breakdown by State";
proc freq data=RESPIR_VARS;
	table State / nocum;
run;

/* barplot for State */
proc sgplot data=RESPIR_VARS;
	vbar State;
run;


**********************************************
/* Counts and relative frequencies for Race */
**********************************************;
title "Breakdown by Marital Status";
proc freq data=RESPIR_VARS;
	table MaritalStatus / nocum;
run;

/* barplot for Marital Status */
proc sgplot data=RESPIR_VARS;
	vbar MaritalStatus;
run;
```

```
********************************************************
/* Counts and relative frequencies for Tumor Behavior */
********************************************************;
title "Breakdown by Tumor Behavior";
proc freq data=RESPIR_VARS;
	table TumorBehavior / nocum;
run;

/* barplot for Tumor Behavior */
proc sgplot data=RESPIR_VARS;
	vbar TumorBehavior;
run;


***************************************************
/* Counts and relative frequencies for Insurance */
***************************************************;
title "Breakdown by Insurance";
proc freq data=RESPIR_VARS;
	table Insurance / nocum;
run;

/* barplot for Insurance */
proc sgplot data=RESPIR_VARS;
	vbar Insurance;
run;
```