Descriptive stats for Data with Shortage

	Year of diagnosis	Histologic Type ICD-O-3	Survival months	Patient ID	Regional nodes examined (1988+)	Regional nodes positive (1988+)	Age at diagnosis	County	Year of follow-up recode	Begin Calc Age (Adjusted)
count	294397.000000	294397.00000	294397.000000	2.943970e+05	294397.000000	294397.000000	294397.000000	294397.000000	294397.000000	294397.000000
mean	2004.091081	8199.62430	70.045140	2.232567e+07	13.701733	31.507098	66.470956	58.765993	2009.793422	66.470956
std	7.138460	158.82469	70.748564	1.205661e+07	19.887476	44.666274	14.172602	50.560803	7.158334	14.172602
min	1992.000000	8000.00000	0.000000	5.400000e+01	0.000000	0.000000	0.000000	1.000000	1992.000000	0.000000
25%	1998.000000	8140.00000	13.000000	1.237613e+07	0.000000	0.000000	56.000000	27.000000	2004.000000	56.000000
50%	2004.000000	8140.00000	45.000000	2.211828e+07	9.000000	2.000000	67.000000	37.000000	2013.000000	67.000000
75%	2010.000000	8240.00000	110.000000	3.342199e+07	17.000000	98.000000	77.000000	85.000000	2016.000000	77.000000
max	2016.000000	9971.00000	300.000000	4.154140e+07	99.000000	99.000000	99.000000	303.000000	2016.000000	99.000000

Descriptive stats for data without shortage

	Year of diagnosis	Histologic Type ICD-O-3	Survival months	Patient ID	Regional nodes examined (1988+)	Regional nodes positive (1988+)	Age at diagnosis	County	Year of follow-up recode	Begin Calc Age (Adjusted)	Number of Intervals (Calculated)
count	5998.000000	5998.000000	5998.000000	5.998000e+03	5998.000000	5998.000000	5998.000000	5998.000000	5998.000000	5998.000000	5998.000000
mean	1998.861621	8184.571691	96.497332	1.978205e+07	13.692231	29.923474	66.485162	763.717573	2006.800767	66.485162	97.406302
std	6.310079	135.812756	89.810199	6.255389e+06	20.186061	43.851714	13.428273	316.180971	8.155207	13.428273	89.726833
min	1992.000000	8000.000000	0.000000	1.560074e+07	0.000000	0.000000	12.000000	28.000000	1992.000000	12.000000	1.000000
25%	1994.000000	8140.000000	18.000000	1.604335e+07	0.000000	0.000000	57.000000	900.000000	1999.000000	57.000000	19.000000
50%	1997.000000	8140.000000	64.000000	1.607377e+07	9.000000	2.000000	68.000000	912.000000	2007.000000	68.000000	65.000000
75%	2000.000000	8210.000000	172.750000	2.012158e+07	17.000000	98.000000	76.000000	912.000000	2016.000000	76.000000	173.000000
max	2016.000000	9717.000000	299.000000	3.242053e+07	99.000000	99.000000	98.000000	999.000000	2016.000000	98.000000	300.000000

Comparing Two means from both groups from **Survival months:**

With shortage - Mean = 70.045140, SD = 70.748564, n = 294397
Without Shortage- Mean = 96.497332, SD= 89.810199, n = 5998

Best point estimate = Difference in the sample mean is 26.452192
Interpretation - We estimate the mean survival month for the areas without health care shortage was 26.452192 higher than for the areas with health care shortage

Two sample T-test :
Null hypothesis: Area with healthcare shortage and area without healthcare shortage means are same
Alternative: Means are different
Assuming variance is not the same (as sd are not different for both.
t= -22.66788625, p-value = 2.52770576e-109 < 0.001

So we reject the numm hypothesis, means are different.