Benjamin Laufer

Ecornell Data and Visualization Final Project

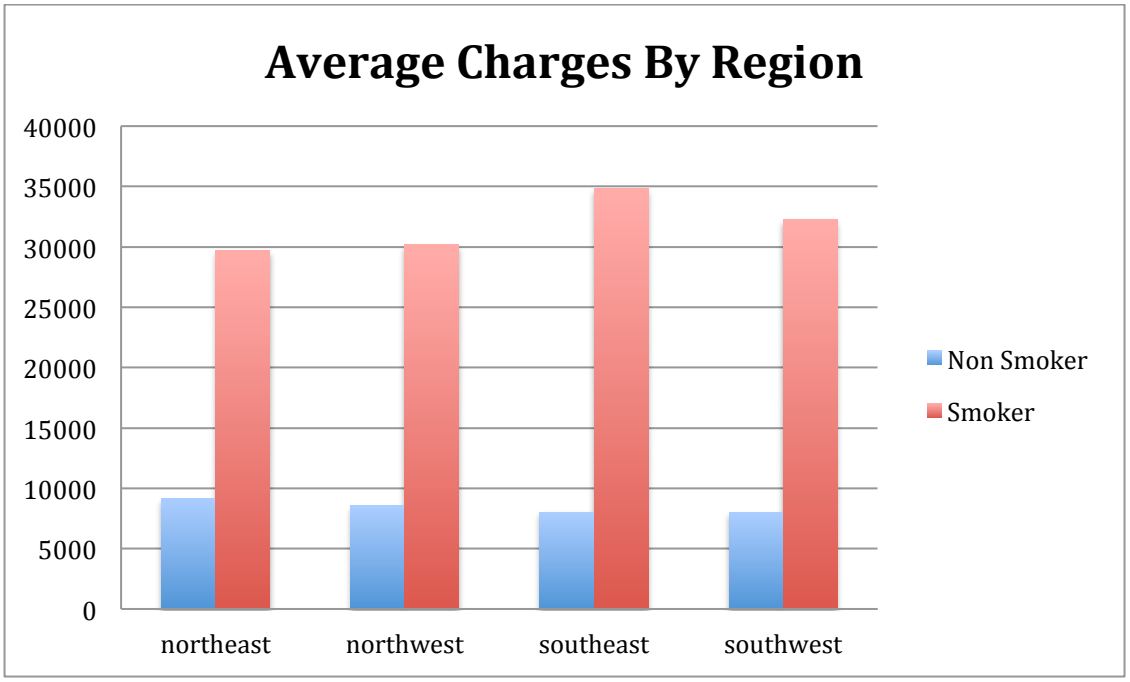
Part 1

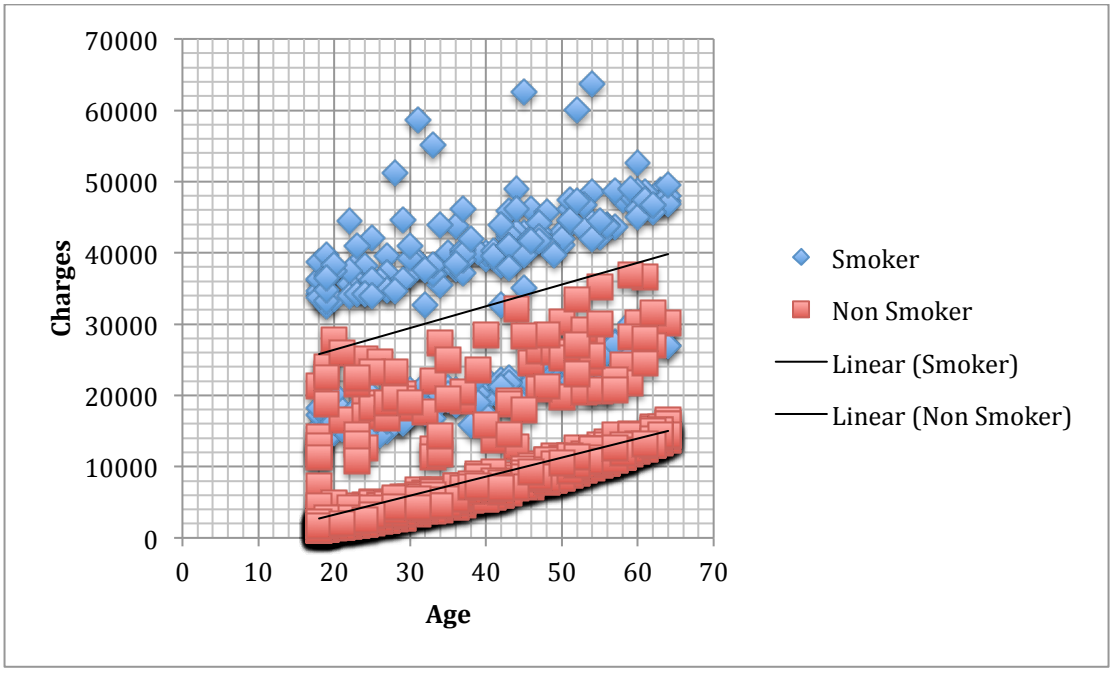
I am examining the affect of smoking on medical charges to help insurance companies better assess what to charge policy-holders. My hypothesis is that smokers spend significantly more on medical costs than non-smokers. The data will be acquired through insurance company looking at new policy-holders and what they were charged in first year. The KPIs I will focus on are the average and standard deviation of what smokers spend on medical bills compared to non-smokers. Factor that will be considered are age, body mass index(BMI) and dependents as I believe all of those will have an effect on medical costs. Considering these factors will ensure that my data is not skewed because of them. The insurance companies will observe new policy-holders and collect the charges over a course the first year. Then I will examine the data to determine if smokers and non-smokers have higher medical costs and if so how great the impact of smoking on charges. To control the sample and make sure it is representative of the population I will select just over 300 people from each region. In order to mitigate bias it is important to consider the factors I mentioned above as that people may lie about whether they are smokers or not

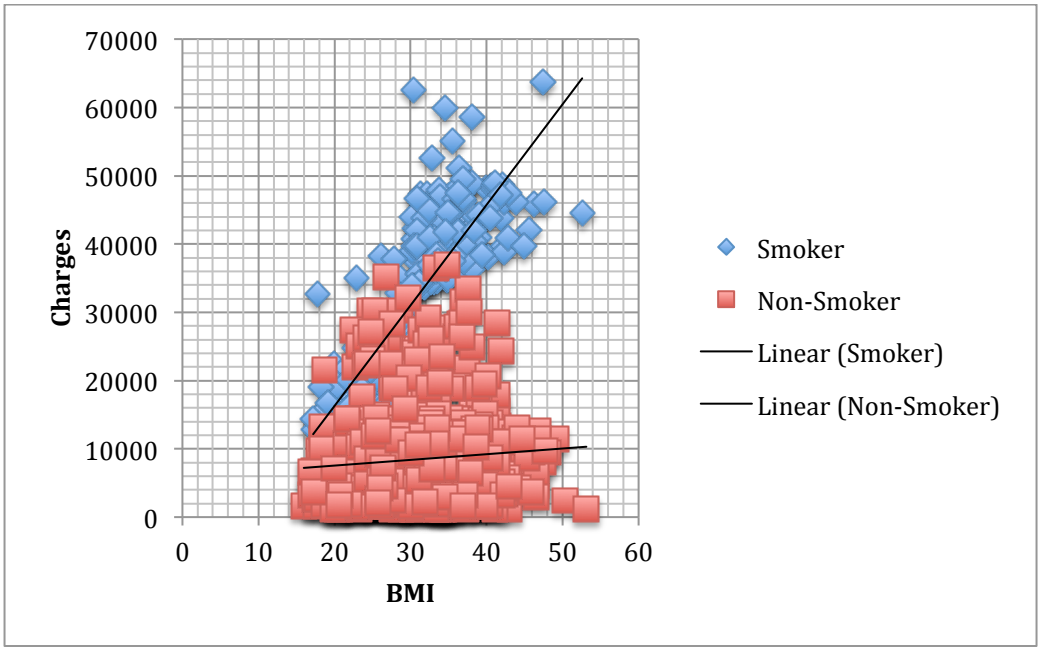
Parts 2 and 3

The summary statistics I will mainly focus on are average and standard deviation of medical charges between smokers and non-smokers. I believe it be suffice and resilient to outliers because of y large sample size

I will use bar charts comparing average charges of smokers to non-smokers, scatter plot graphs showing data on charges to smokers based on weight and age.


Average Charges By Region
40000
35000
30000
25000
20000
15000
10000
5000
0
northeast
northwest
southeast
southwest
Non Smoker
Smoker


70000
60000
50000
40000
30000
20000
10000
0
Charges
0
10
20
30
40
50
60
70
Age
Smoker
Non Smoker
Linear (Smoker)
Linear (Non Smoker)



	Avergage	SD	Median	Max	Min
Smoker	32050.23183	11541.54718	34456.34845	63770.42801	12829.4551
Non-Smoker	8434.268298	5993.781819	7345.4053	36910.60803	1121.8739

The chart above shows that on average smokers are charged over $23,000 more than non-smokers which is about 3.8 times more than what non-smokers pay. The difference was even greater between the medians at over $27,000. The standard deviation in charges was significantly greater for smokers than non-smokers and their min and max were both significantly higher.

The graphs above strongly display that smokers pay way more than non-smokers fro Insurance causes. Interestingly, the correlation is about the same for age of smokers and non-smokers. But there is a very strong increase in correlation in medical charges between smokers and BMI. My data does not take into account previous medical conditions or if people lied about being a smoker but the data is nonetheless conclusive that smokers pay way more than non-smokers in medical costs.

*My data for this project was pulled from Kaggle and the original file was entitled Insurance.csv

children	(All)
sex	(All)

	Column Labels								
	no			yes			Total Average of age	Total Average of charges	Total StdDev of charges
Row Labels	Average of age	Average of charges	StdDev of charges	Average of age	Average of charges	StdDev of charges			
northeast	39.53696498	9165.531672	6157.196439	38.23880597	29673.53647	11589.55977	39.26851852	13406.38452	11255.80307
northwest	39.16853933	8556.463715	6128.553819	39.32758621	30192.00318	11413.8277	39.19692308	12417.57537	11072.27693
southeast	38.67399267	8032.216309	6137.329758	39.73626374	34844.99682	11324.76654	38.93956044	14735.41144	13971.09859
southwest	40.1835206	8019.284513	5489.514421	36.10344828	32269.06349	11270.27959	39.45538462	12346.93738	11557.1791
Grand Total	**39.38533835**	**8434.268298**	**5993.781819**	**38.51459854**	**32050.23183**	**11541.54718**	**39.20702541**	**13270.42227**	**12110.01124**