Final Project

Team 3

IST 687 Fall 2022

Data Set Content

- Variables
  - Age
  - Sex
  - BMI
  - Children
  - Smoker
  - Region
  - Charges
- Data Source - Medical Cost Personal Datasets | Kaggle
- Observations – 1338

Original Business Questions

- What are the demographics of our patient population?
- What variables have the most effect on insurance charges?
- How can we predict an individual's insurance cost based on their variables/ how accurately can we forecast costs for each patient by looking at the data?
- What variables in the dataset can an individual control to potentially reduce their insurance charges, and by how much?
- Other than behavior of individuals such as smoking, what can cause increase of medical cost?
- What variables have the strongest correlation?

Demographics

Average Age: 39.21 years

- Range: 18-64

Average BMI: 30.66

- Range: 15.96-53.13

Average Children: 1.09

- Range: 0-5

Average Charges: $13,270.42

- Range: $1,121.87-$63,770.43

Gender:

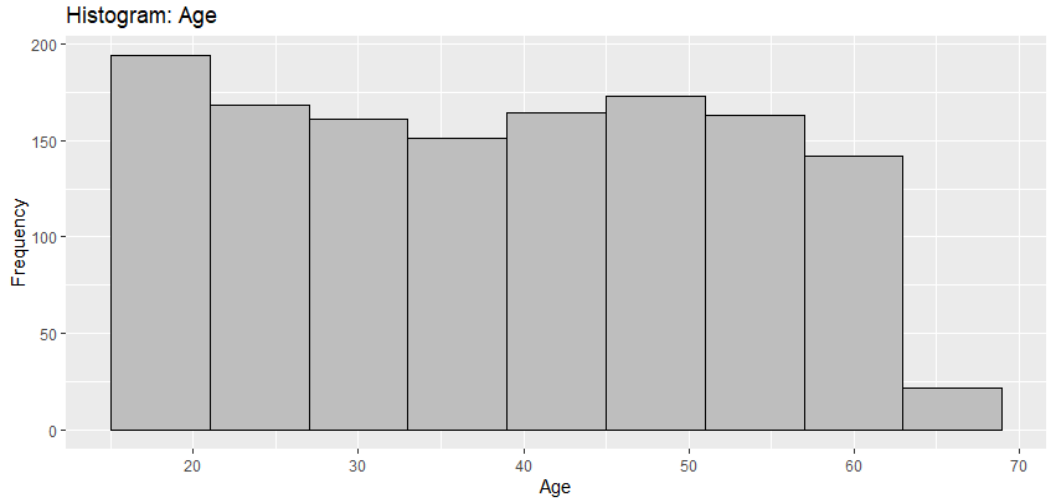
- Female - 48.48%
- Male – 50.52%

Percentage Smokers:

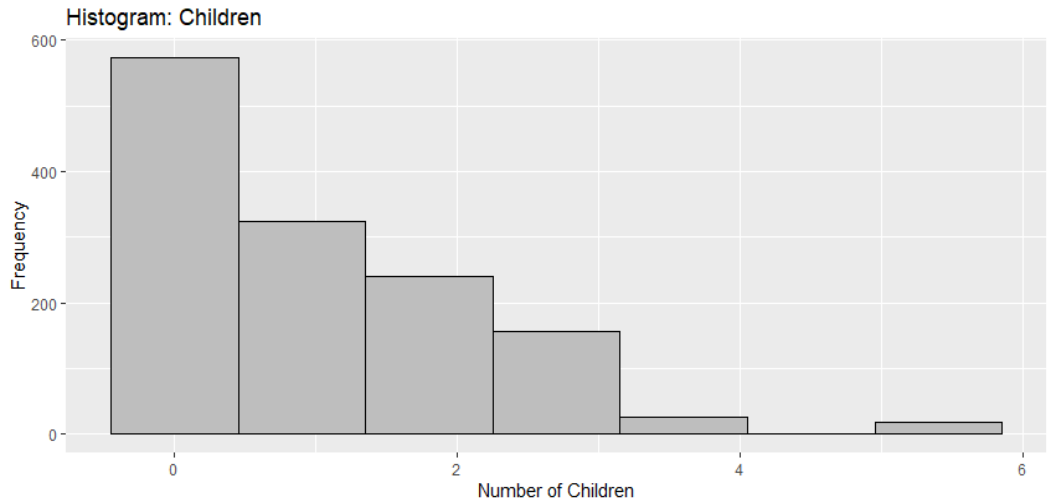
- Smoker – 20.48%
- Non-smoker – 79.52%

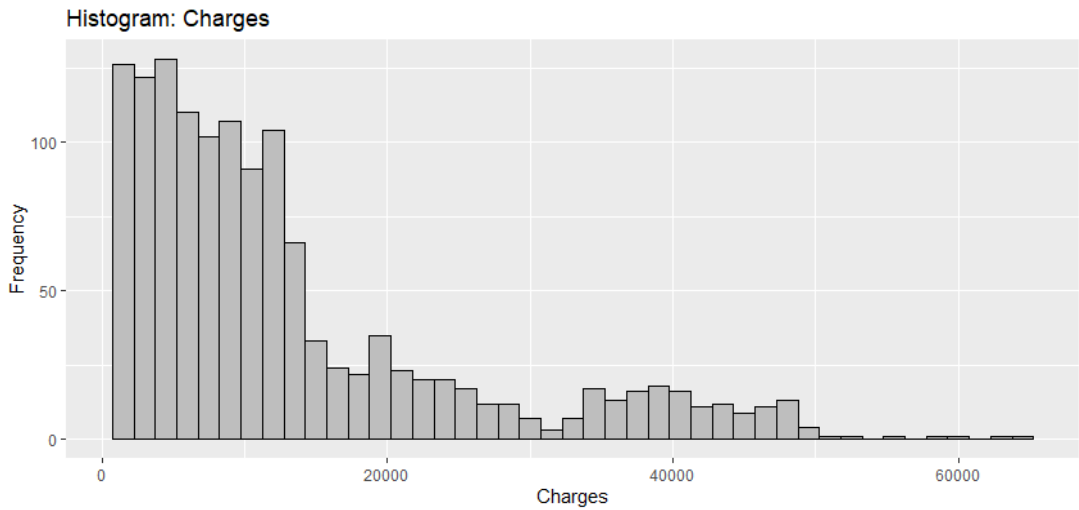
Population by Region:

- Southwest – 325
- Southeast – 364
- Northwest – 325
- Northeast – 324


Histogram: Age
Frequency
200
150
100
50
0
20
30
40
50
60
70
Age


Histogram: BMI
Frequency
600
400
200
0
10
20
30
40
50
60
Body Mass Index


Histogram: Children
Frequency
600
400
200
0
0
2
4
6
Number of Children


Histogram: Charges
Frequency
100
50
0
0
20000
40000
60000
Charges

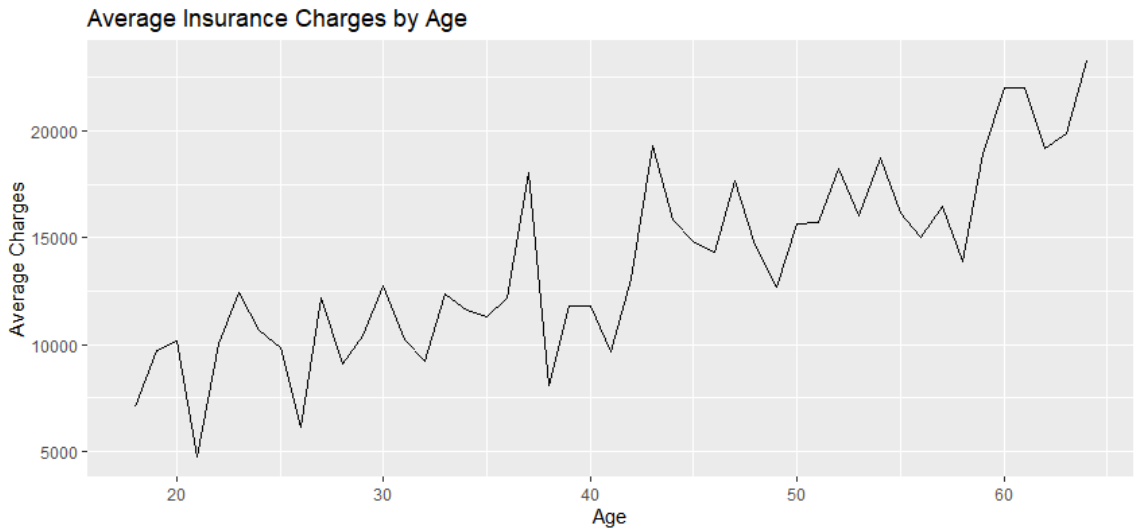
Multivariate Linear Regression Model

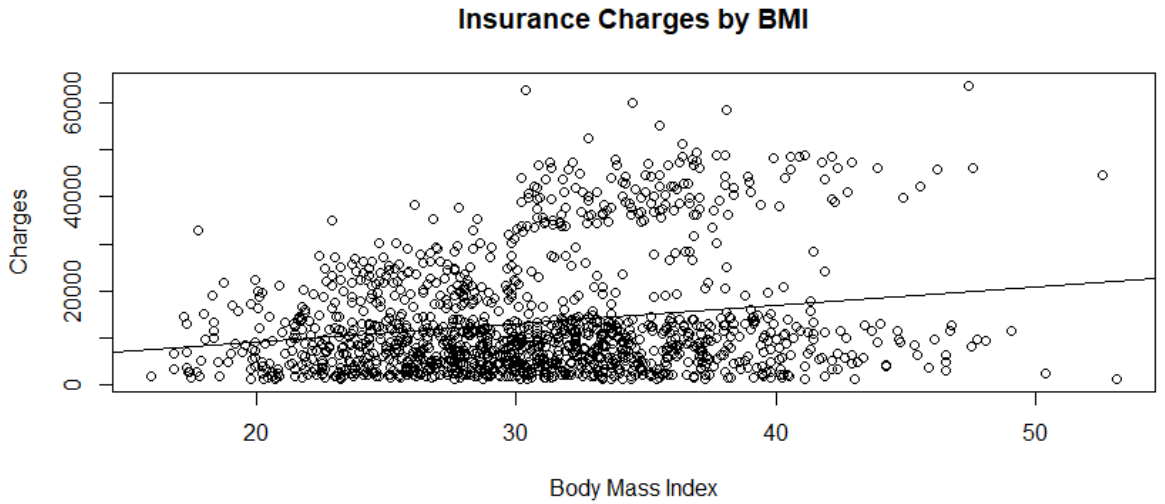
- Initial Regression Model
  - charges = age + bmi + children + female + smoker + southwest + southeast + northwest
  - F-statistic = < 2.2e-16
  - Adjusted R-squared: 74.94%
- Statistically Significant Model
  - charges = age + bmi + children + smoker
  - F-statistic = <2.2e-16
  - Adjusted R-squared = 74.89%

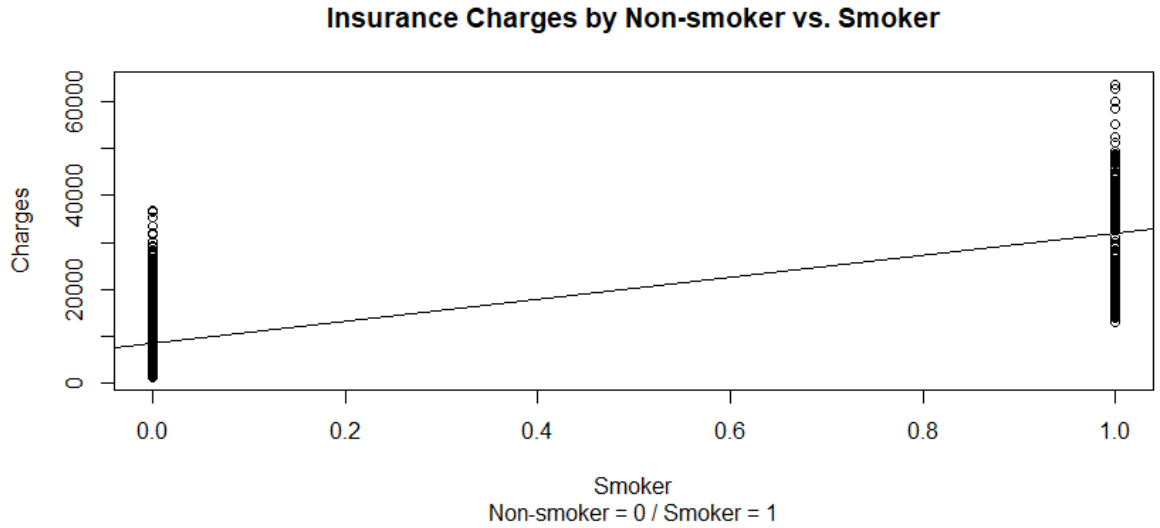
MLR Model Drilled Down

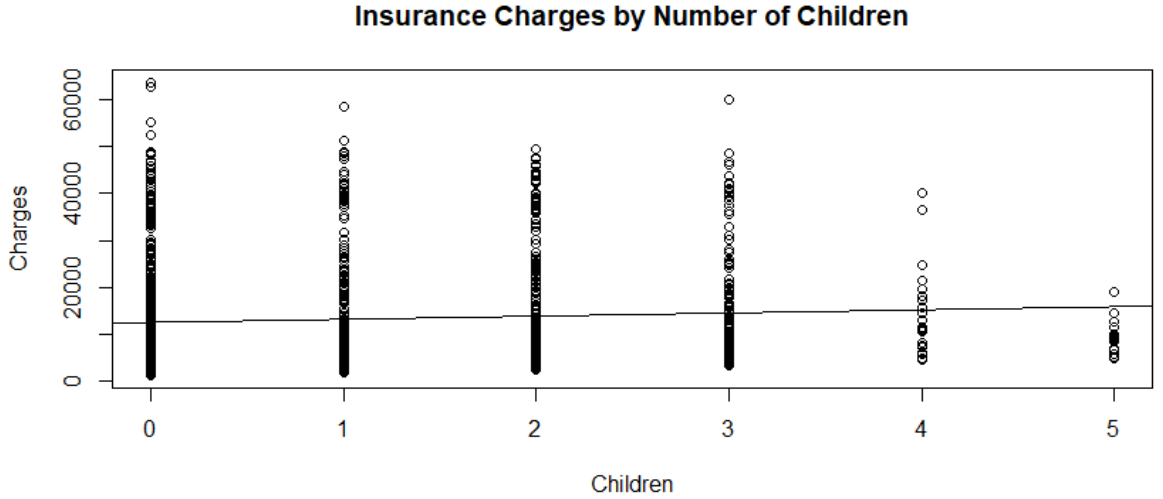
- Remove Age
  - charges = bmi + children + smoker
  - Adjusted R-squared: 66.07%
- Remove BMI
  - charges = age + children + smoker
  - Adjusted R-squared: 72.31%
- Remove Children
  - charges = age + bmi + smoker
  - Adjusted R-squared: 74.69%
- Remove Smoker
  - charges = age + bmi + children
  - Adjusted R-squared: 11.81%

Significant Variables Effects on Charges



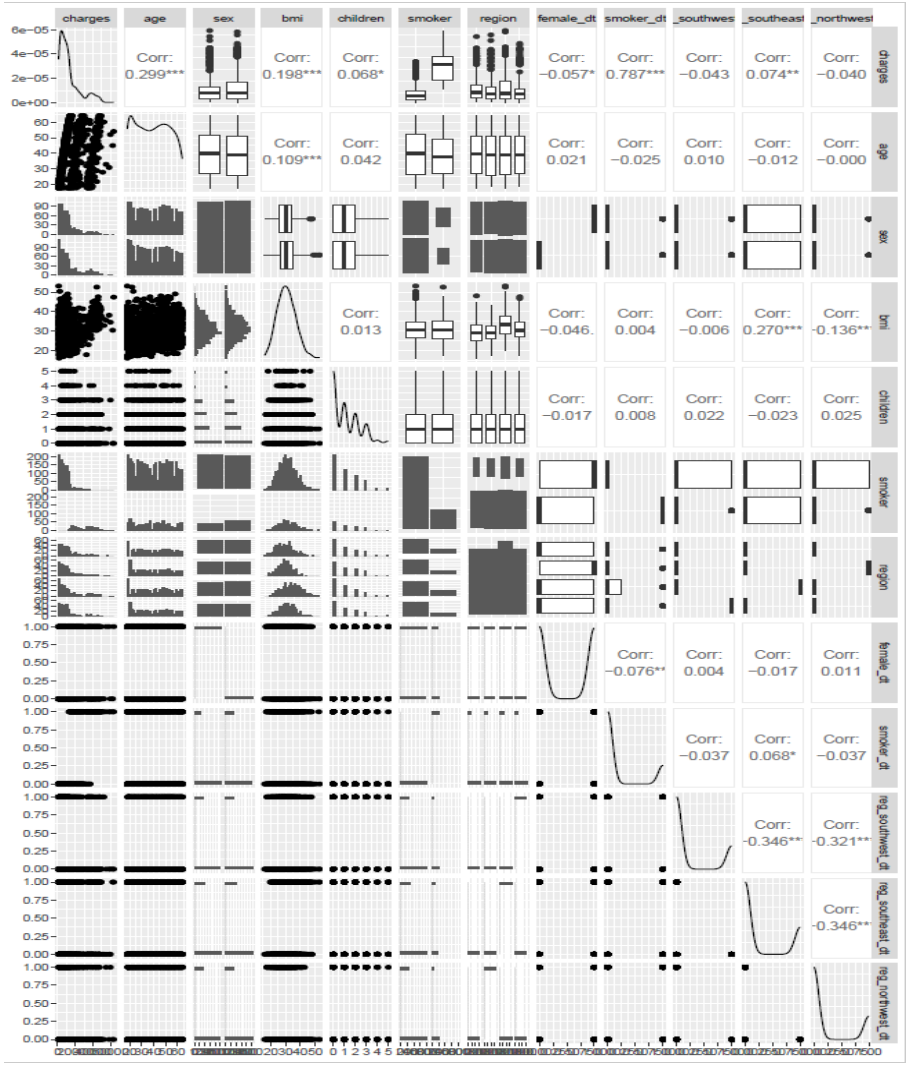






Correlation of Variables

- Most significant correlation
  - Charges and Smoker: 78%
  - Charges and Age: 30%
  - Southest Region and BMI: 27%



Prediction Model/Function for Business App

```r
charges_function <- function(age,bmi,children,smoker) {
  model_charges <- lm(charges ~ age + bmi+ children + smoker_dt, data = ins_df)
  pred_df <- data.frame(age = age, bmi = bmi, children = children, smoker_dt = smoker)
  pred_charges <- predict(model_charges,pred_df, type = "response")
  return(pred_charges)
}
charges_function(27,32,1,0)
```

Output: $5,631.92

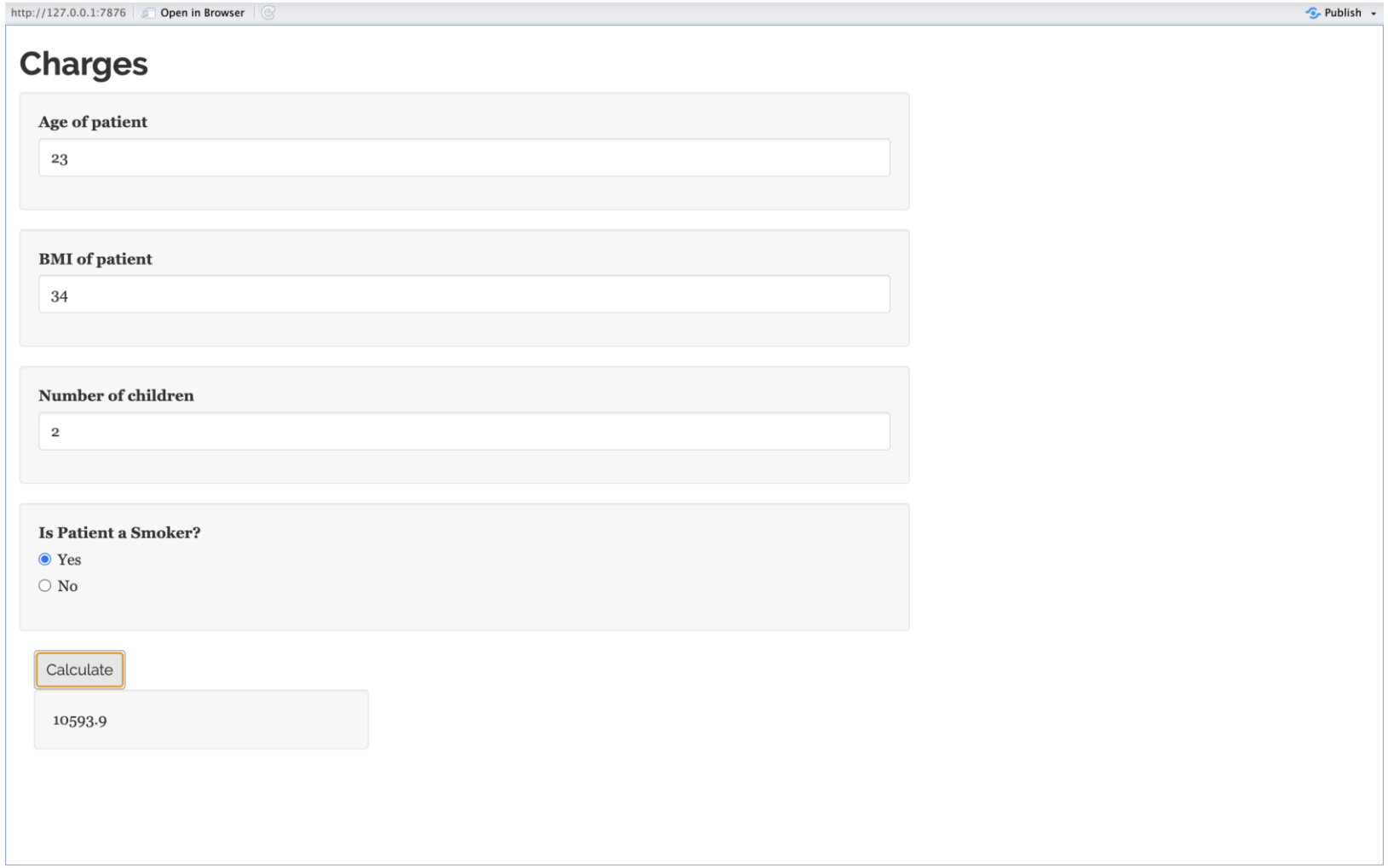
Additional Business Questions

- Controllable Variables
  - BMI
  - Smoking
  - Number of Children*
- Uncontrollable Variables
  - Age

R-squared: 74.89%	F-Statistic: <2.2e-16	
Variable	**Coefficient**	P-Value
Intercept	-12102.77	< 2e-16
Age	257.85	< 2e-16
BMI	321.85	< 2e-16
Children	473.50	0.000608
Smoker	23811.40	< 2e-16

*Assumes children aren't already born

Shiny App for Web User



Limitatations of the Dataset

- Several large outliers that may affect assumptions made in the linear regression analysis
- Inabilty to address the high average cost of healthcare
- Dataset is from one insurance company
- Prediction models failing due to global events that disrupt healthcare system, e.g. the COVID-19 pandemic

Future Topics to Address The Limitations

- Dataset from various insurance companies and cross analysis of the cost across healthcare systems
- Dataset from other sources including government sources on the number of unisured patients and the role they play in the cost of healthcare
- Transparancy in medical bills including administrative cost and hidden fees to address discrepancies in the cost