

# Life Insurance Cost Analysis Project



U of T Data Analytics Bootcamp - Group 9

Fayyaz Hannan  
Moataz Yaakoub  
Siqi Ou  
Junior Vantuil

# Table of Contents

- Project Overview
- Technology
- Database
- Machine Learning
- Visualization
- Real World Application

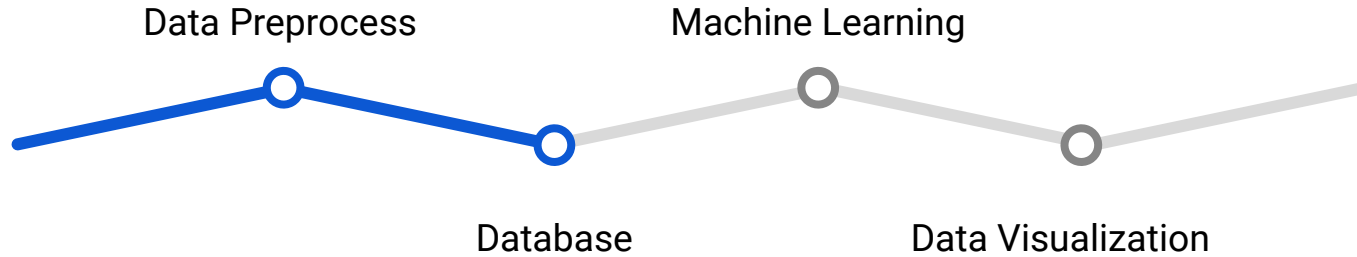
---

# Project Overview

Life insurance is a valuable tool for Canadians to protect themselves from financial difficulties in the case of an unfortunate event, but many people are turned away from the cost. It would be beneficial to identify the variables related to life insurance premiums and guide customers to purchases.

We conducted analysis on a set of real life demographics data including age, gender, income, life insurance rates, and used machine learning to estimate the cost of life insurance and make recommendations to clients.

# Technology

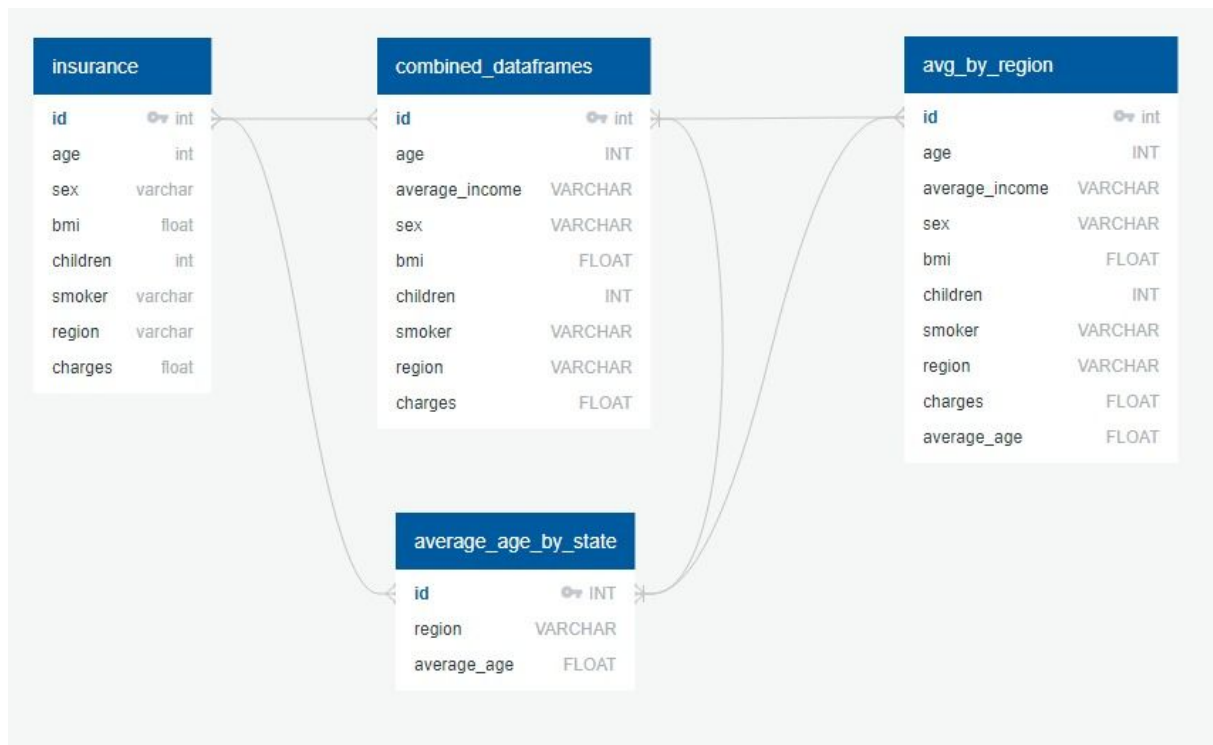


# Data Source

1. 1338 customer demographics data collected from Kaggle
2. Average income percentile by age collected from DQYDJ
3. List of US states by life expectancy collected from Wikipedia

	id	age	average_income	sex	bmi	children	smoker	region	charges	average_age
0	0	18	10753.5	male	33.77	1	no	southeast	1725.5523	78.43
1	1	18	10753.5	male	34.10	0	no	southeast	1137.0110	78.43
2	5	18	10753.5	male	31.68	2	yes	southeast	34303.1672	78.43
3	9	18	10753.5	female	36.85	0	yes	southeast	36149.4835	78.43
4	11	18	10753.5	female	38.28	0	no	southeast	1631.8212	78.43

# Database - ERD



# Data Processing - ETL

- Reformat and load in the raw data into a dataframe in Jupyter Notebook using Python Pandas Library
- Combined 3 different dataset by inner join: demographics data, average income, and life expectancy by states. "age" was the keyword used to combine the dataframes.
- The merged CSV files are connected to PostgreSQL Database, data is consolidated and stored on AWS
- 2 tables are created in SQL, inner join applied
- Data is scaled and normalized in the model

# Machine Learning Model

Questions we hope to answer with the data

**ML Goal:** For the machine learning analysis, we used supervised machine learning model to explore the relationships between demographics variables and cost of life insurance to predict cost of life insurance.

**Implications:** By finding key elements that had strong correlations to the cost, such as age and income, we can provide a strong tool for customers who are considering purchasing life insurance.

## Demographics Data explored:

- Age
- Gender
- BMI
- Number of Children
- Smoker (Yes or No)



# Feature Engineering

- Target: identify which variable is highly correlated to the insurance cost
- Features: demographics data including age, bmi, number of children, average income, etc.
- Feature Engineering: use python to perform exploratory data analysis for identification, dropped uncorellated variables - sex and region
- Decision making: determine which variable is in high correlations and continue with training and testing, in our case smoker/non smoker had big impact on the life insurance costs

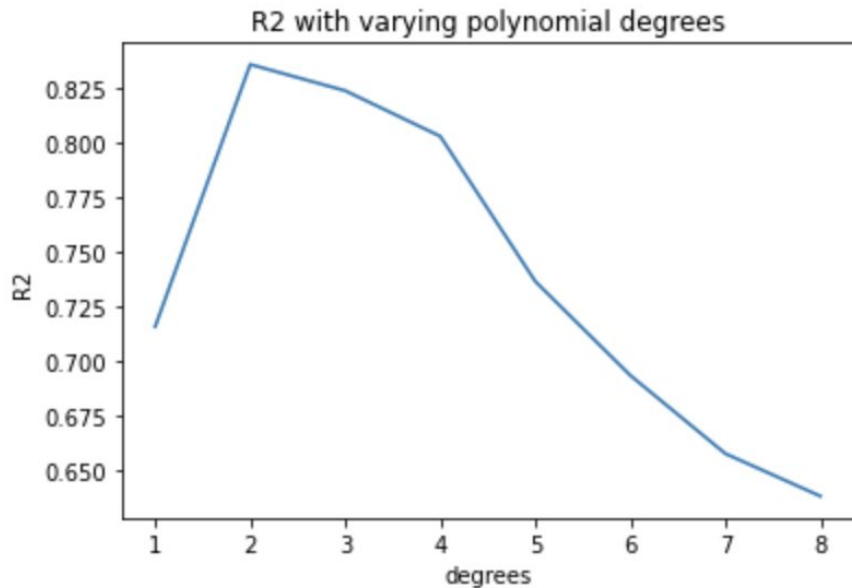
```
dummies = pd.get_dummies(complete_data_df[['sex', 'smoker', 'region']], drop_first=True)
text_features = pd.concat([complete_data_df.drop(['sex', 'smoker', 'region'],axis=1), dummies],axis=1)
text_features.head()
```

	id	age	average_income	bmi	children	charges	average_age	sex_male	smoker_yes	region_northwest	region_southeast	region_southwest
0	0	18	10753.5	33.77	1	1725.5523	78.43	1	0	0	1	0
1	1	18	10753.5	34.10	0	1137.0110	78.43	1	0	0	1	0
2	5	18	10753.5	31.68	2	34303.1672	78.43	1	1	0	1	0
3	9	18	10753.5	36.85	0	36149.4835	78.43	0	1	0	1	0
4	11	18	10753.5	38.28	0	1631.8212	78.43	0	0	0	1	0

# Machine Learning Model - Polynomial Regression

**Benefits:** Great for determining relationship between multiple inputs (independent variables) and single output (dependent variable)

**Limitations:** Very sensitive to outliers. The model will not perform as expected if there are outliers. This means the data we input into the model must be very clean

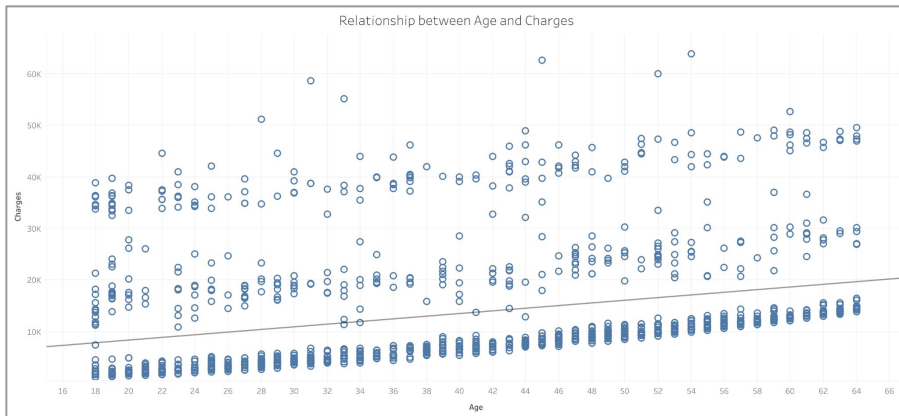


	Actual	Predicted
1	1137.01100	8785.042078
266	9549.56510	13655.922585
1145	4320.41085	7567.592284
1148	5125.21570	6901.977921
760	33900.65300	18640.067289
...	...	...
1318	14119.62000	18981.727525
1134	21472.47880	29179.698731
155	5245.22690	3642.861827
838	15820.69900	9509.292537
606	44641.19740	37874.131071

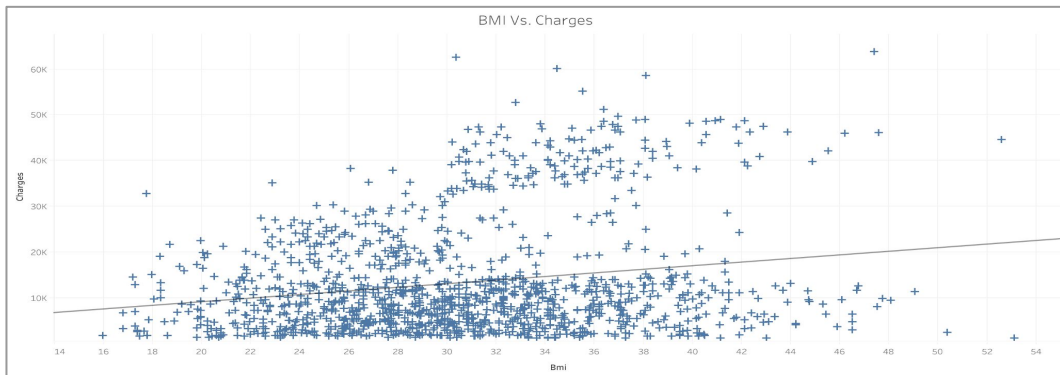
268 rows x 2 columns

# Dashboard

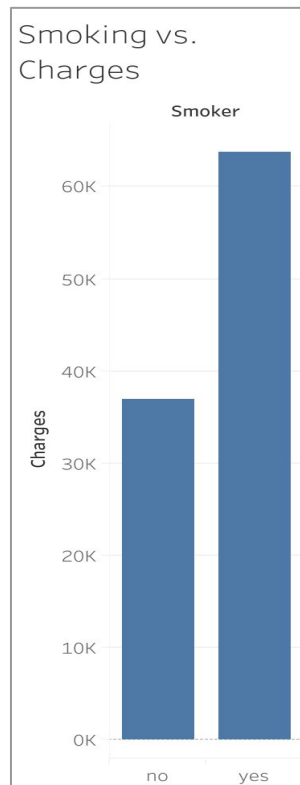
Age vs. Charges



BMI vs. Charges



Smoking vs. Charges



# Interactive Webpage

Link of webpage: <https://life-insurance-quote-project.herokuapp.com/predict>

## Life Insurance Quote

Try our App!! Get an estimated life insurance quote with a few clicks!!

Want to learn how we created the following model?

[CLICK HERE TO LEARN](#)

## Superior Collaboration Visualize Quality

Why is life insurance important? Buying life insurance protects your spouse and children from the potentially devastating financial loss that can occur in the event of an untimely death. Life insurance provides financial security, helps to pay off debts, helps to pay living expenses, and helps to pay any medical or final expenses.

## Life Insurance Prediction

Age  income  bmi  children    
smoker\_yes

8392.641076445612

# Real World Application

By applying this model, we could determine which clients may be at risk, and which clients could be offered premium rates:

## For Clients:

- Quick quote calculation by simply entering demographics info
- Find the right coverage for each individual's unique situation
- Plan ahead and protect loved ones

## For Companies:

- Ready to use machine learning model for application
- Identify premium rates for clients by only a few client data inputs

© Randy Glasbergen  
www.glasbergen.com



**“If I let myself get bitten by a vampire so I become immortal and only a wooden stake can kill me, can I get a better rate on my life insurance?”**

# Key Takeaways

**Challenges:** Needed more variables to test the model, we tried to find different statistics and combine data frames, but it would be good if we have an original dataset that includes more columns.

**Recommendation for future analysis:** Look into different life insurance costs by country and region. The dataset we had was US demographics, if we could find any Canadian data and the life insurance companies information, it would be great for companies to compare and identify their competitive advantages.

**What could have done differently:** Do more research to find datasets which includes actual age data instead of scraped average numbers into our Machine Learning Model as we believe it might improve the accuracy of model.

THANK YOU

# Dashboard Storyboard Outline

By exploring and processing different sets of real life demographics data, we are going to create an insurance cost estimating tool for:

1. Customers who wanted to know their life insurance costs
2. Advisors who are helping clients to identify their status and which rates to offer

This will be presented on a webpage where users can input their demographics data to calculate life insurance premiums, along with data findings presented in several charts.

## **Tools used:**

Javascript, CSS, HTML

## **Interactive elements on webpage:**

An input box for age, income, bmi, # of children, and Smoker or Non Smoker, then it will return the predicted rates