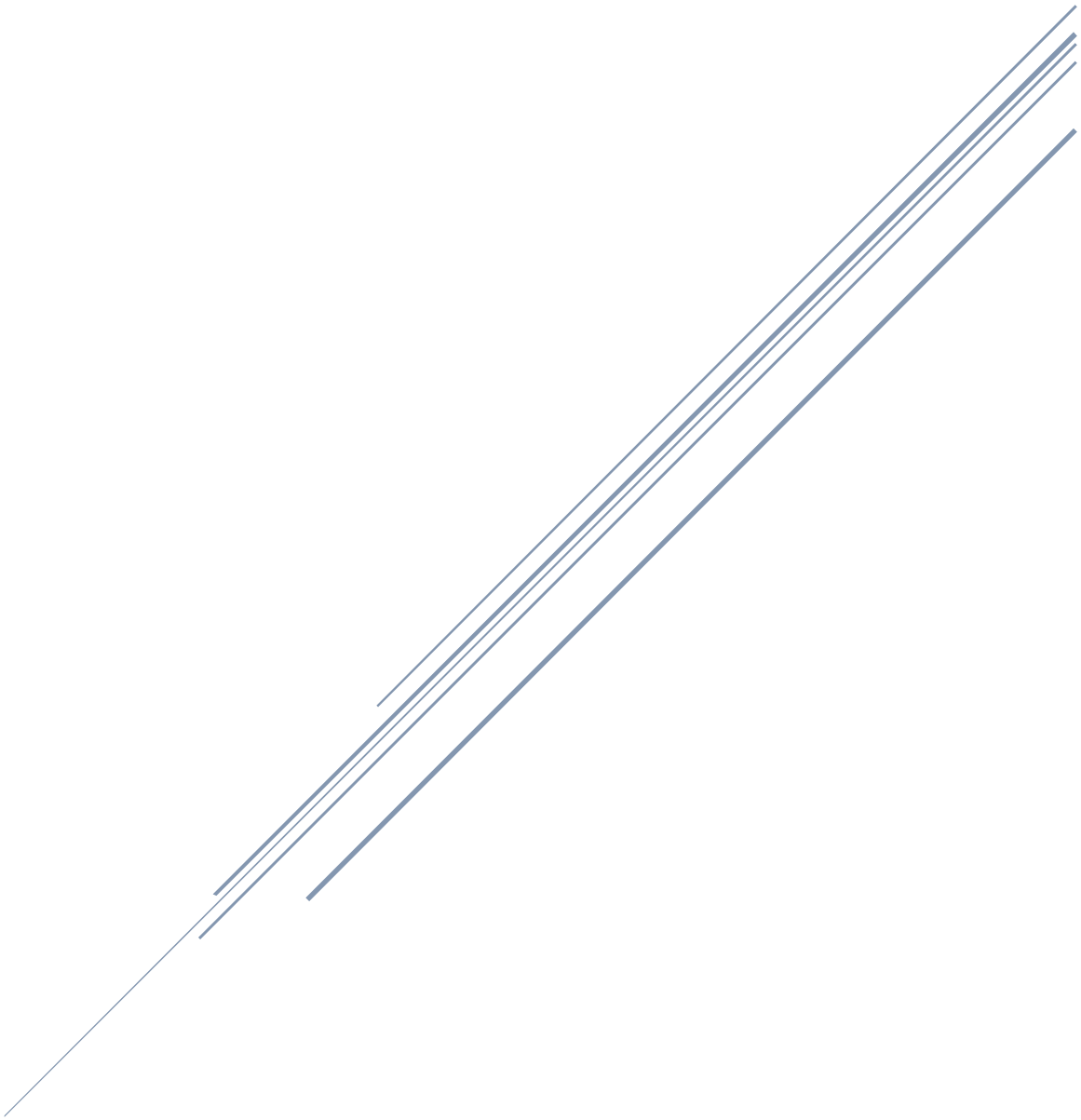
LINEAR-REGRESSION PROJECT PROPOSAL

SDAIA T5 Data scince boot camp

Norah Alqahtani, Batoul Alosaimi, and Shroaq Almutairi



Introduction:

As the lead data scientist at NSB company we've had an opportunity to work with ACME Insurance Inc. that offers affordable health insurance to thousands of customers all over the United States., we're tasked with creating an automated system to estimate the annual medical expenditure for new customers, using information such as their age, sex, BMI, children, smoking habits, and region of residence.

Estimates from our system will be used to determine the annual insurance premium (amount paid every month) offered to the customer. Due to regulatory requirements, we will explain why our system outputs a certain prediction.

From this web page https://gist.github.com/meperezcuello/82a9f1c1c473d6585e750ad2e3c05a41 we're going to gather data using beautiful soup and selenium in python to be able after that to apply linear regression to help us predicting the annual medical expenditure for new customers.

Question/need:

1. are any relevant corralection between the ratio of multiple quantitative variables; such as age/bmi vs charges/children?
2. which are the top 10 states with the highest charges(based on this sample data)?
3. Can we plot the regions with oldest population and higest charges?

Data Description:

This dataset contains 1300 rows of insured data, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker and Region. The attributes are a mix of numeric and categorical variables.

Attribute Information:

- Age: age of the primary beneficiary.
- Sex: Insurance contractor gender, female / male.
- BMI: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / $m2$) using the ratio of height to weight, ideally 18.5 to 24.9.
- Children: Number of children covered by health insurance / Number of dependents.
- Smoker: smoker / Non - smoker.
- Region: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- Target Column(Charges): Persons Annual Medical charges.

Tools:

- beautiful soup and selenium in python (Scraping the web)
- Pandas and Numpy (Exploring the data)
- Matplotlib and Seaborn (Visualizing the data)
- Sci-kit Learn (linear Regression model)