The University of Memphis
Computer Science
COMP 7/8150-Fundamentals of Data Science
Project Proposal: **Health Insurance Premium at PSI**
Vinay Badam, Nikhil Banka, Data Science, Fall 2022

PROJECT GOAL AND SCOPE

The goal of this project is to predict the health insurance premiums of customers of PSI six months in advance. The solution will be one python program that predicts the premium with an average relative error of at most 20%.

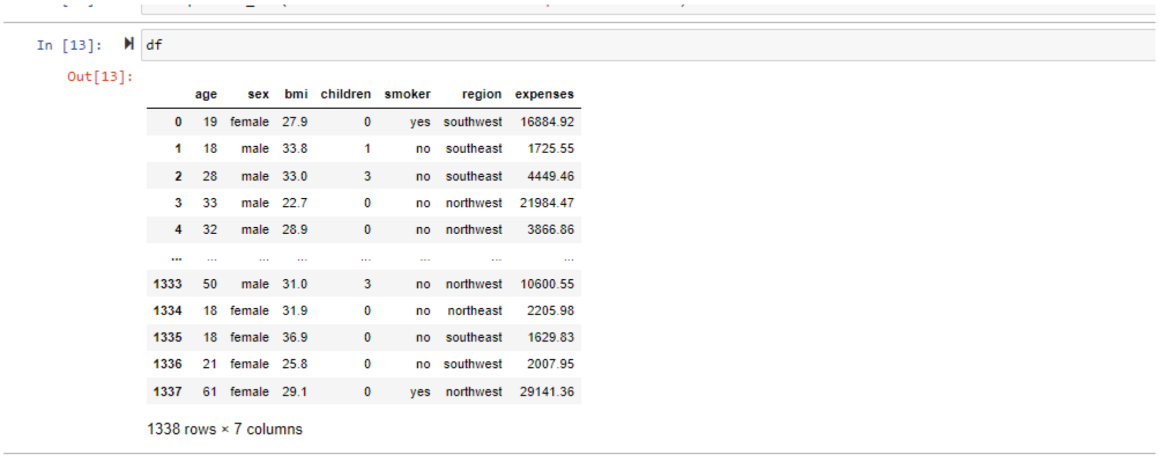
BACKGROUND/LIT REVIEW

{VB} In the event of a major illness or accident, it safeguards your financial situation. One form of policy that helps people budget for and pays for medical costs is health insurance. Insurance companies use a person's age, physical condition, family situation, and past medical bills to predict their future medical costs in order to decide how much to charge for coverage.

The full population under consideration for the project would be clients of the insurance provider PSI. This dataset [2] contains data on the insurer, their dependents, their medical costs over a year, etc. The prediction is based on the dataset's other parameters such as BMI, smoker, age, etc. In the dataset, we can also see the previous expenses paid by clients for the previous term. The dataset [2] contains the following data items:

1. Age
2. Sex
3. BMI: Body Mass Index of the insurer
4. Children: determines if the insurer has any children to be covered.
5. Smoker: Determines whether he/she is a smoker or non-smoker
6. Region: Residence of the insurer
7. Charges: Yearly medical charge

The dataset contains 1338 records and has a dimensionality of 7

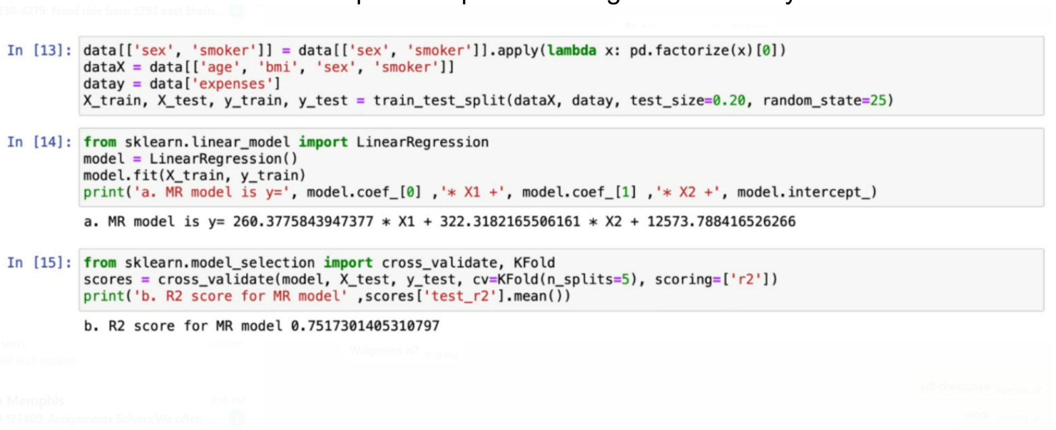


```
In [13]:  df
Out[13]:
```

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86
...	...	...	...	...	...	...	...
1333	50	male	31.0	3	no	northwest	10600.55
1334	18	female	31.9	0	no	northeast	2205.98
1335	18	female	36.9	0	no	southeast	1629.83
1336	21	female	25.8	0	no	southwest	2007.95
1337	61	female	29.1	0	yes	northwest	29141.36

1338 rows × 7 columns

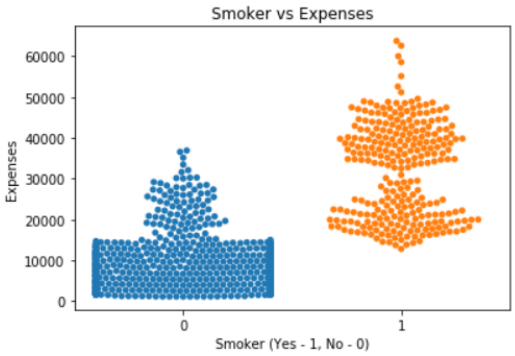
{NB} The articles [3] and [4] worked on the insurance prediction problem using linear regression based on the datasets [2] and data available from the dataset [2] having required data of insurance firm clients which resulted in a relative error or mean squared error for the data to be higher than expected. The author [4] divided the features into numerical and nominal features. These nominal features were converted into factors with a numerical value. The author first, to understand the data plotted the dispersion graphs of the features and then calculated the mean and median for age, BMI, and expenses. The author found that there is a very slight variation in the mean and median for age and BMI but for expenses, there is a high variation. The author then has done a scatter plot between age and expenses with respect to BMI to understand why people with low age have high BMI and high age with low BMI. So, the three features, age, BMI, and smoker have a relationship with expenses and give satisfactory MR value.



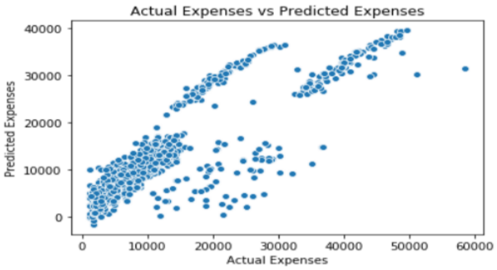
```
In [13]: data[['sex', 'smoker']] = data[['sex', 'smoker']].apply(lambda x: pd.factorize(x)[0])
         dataX = data[['age', 'bmi', 'sex', 'smoker']]
         datay = data['expenses']
         X_train, X_test, y_train, y_test = train_test_split(dataX, datay, test_size=0.20, random_state=25)

In [14]: from sklearn.linear_model import LinearRegression
         model = LinearRegression()
         model.fit(X_train, y_train)
         print('a. MR model is y=', model.coef_[0] ,'* X1 +', model.coef_[1] ,'* X2 +', model.intercept_)

         a. MR model is y= 260.3775843947377 * X1 + 322.3182165506161 * X2 + 12573.788416526266

In [15]: from sklearn.model_selection import cross_validate, KFold
         scores = cross_validate(model, X_test, y_test, cv=KFold(n_splits=5), scoring=['r2'])
         print('b. R2 score for MR model' ,scores['test_r2'].mean())

         b. R2 score for MR model 0.7517301405310797
```

CASE STUDY.

Venkat Murali [4] presented a model for insurance premium prediction [4] using linear regression. Here Linear regression is used to predict the value of a variable based on the value of another variable. The variable we are trying to predict is called the dependent variable which is Expenses and the variable used to predict the other variable's value is called the independent variable which are sex, age, and BMI. Here the author used basic linear regression to predict the insurance premium. The author of this article used a small dataset from the entire dataset and then created a linear regression training model and fed 30% of the data and then plotted a graph to see the actual versus predicted expenses and r squared value of 0.732.

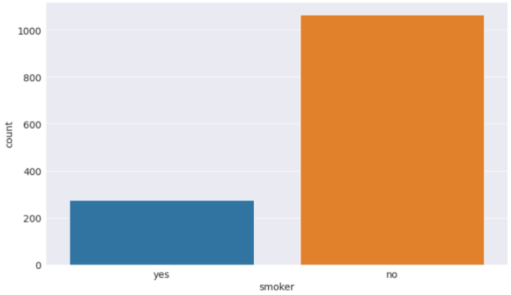


From the above diagram, we can observe that Smokers (yes or no) on X -axis and Expenses on Y-axis, when we plot the graph, we can conclude that expenses are very high for smokers when compared to non-smokers.

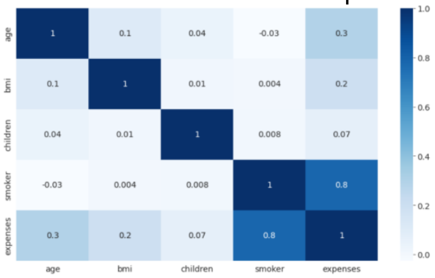


From the above paper, I've considered the entire dataset (1338 records) after analyzing the data and finding the relationship between different features. From the dataset, I first plotted each variable's significance before creating a model.

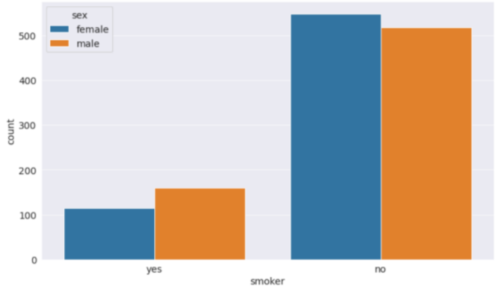
Number of insurers by smoking



Correlation of heat map



Number of insurers by sex



TAKE-HOME DELIVERABLE

The deliverables for this project will be one python program that predicts the premium with an average relative error of at most 20% over the whole population as estimated from the dataset, not just the dataset.

REFERENCES

[1] Centers for Medicare Services (CMS)
https://data.cms.gov/search?keywords=HEALTH%20INSURANCE&sort=Relevancy

[2] https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression

[3] https://www.kaggle.com/datasets/hhs/health-insurance-marketplace

[4] https://www.kaggle.com/datasets/hhsasd/health-insurance-marketplace