

Cost of Health Insurance

By: Kellan Bouwman, Robert Hancock, Andre Salazar





Stage One

What is the Dataset - Personal Medical Cost - [Link](#)

- Personal information for medical reasons -
bmi/gender/age/region/smoker/children/charges

What is the questions/answer you expect to find

- Predict charges

What are the benefits of this project

- Helping people know how much they should expect for personal medical costs



Stage Two

Preparing your dataset:

How you get your data? Pre-made or scrap from the web, or DB? i.e. Kaggle (Links to an external site.)? UCI (Links to an external site.)?

- <https://www.kaggle.com/mirichoi0218/insurance/data>

How you clean your data?

- The data was mostly cleaned as we received it, when through and averaged missing values

How did you prepare it?

- Arranged by two fold for the categorical data smoker and gender

Need to acquire more related dataset?

- Yes, diabetes cost of insulin would also be helpful for the prediction of personal charges



Stage Three

Descriptive analysis

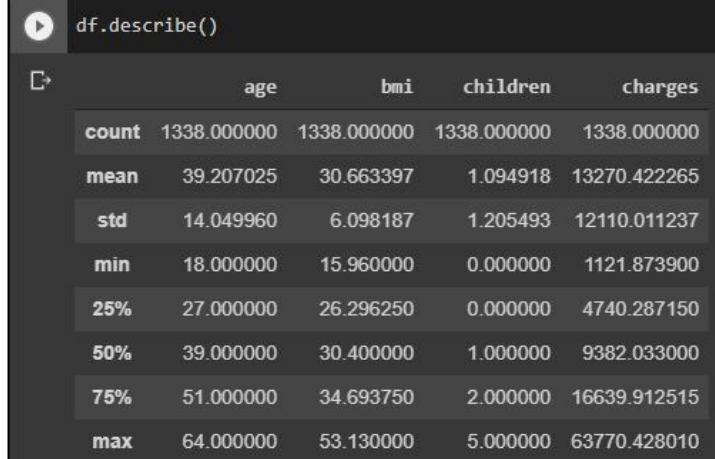
- There is great variance in charges, from the 75% percentile to max charge there is over \$4000, in difference

Get to know data, develop hypotheses, patterns? Anomalies?

- Age and bmi seem to be controlling factor in both genders and smokers

Problem with data? How to improve this dataset in the future?

- Lack of information of which insurance agency they are with, and diabetes as well as insulin costs, and any other preexisting or continuing conditions, and finally what the charges are for.



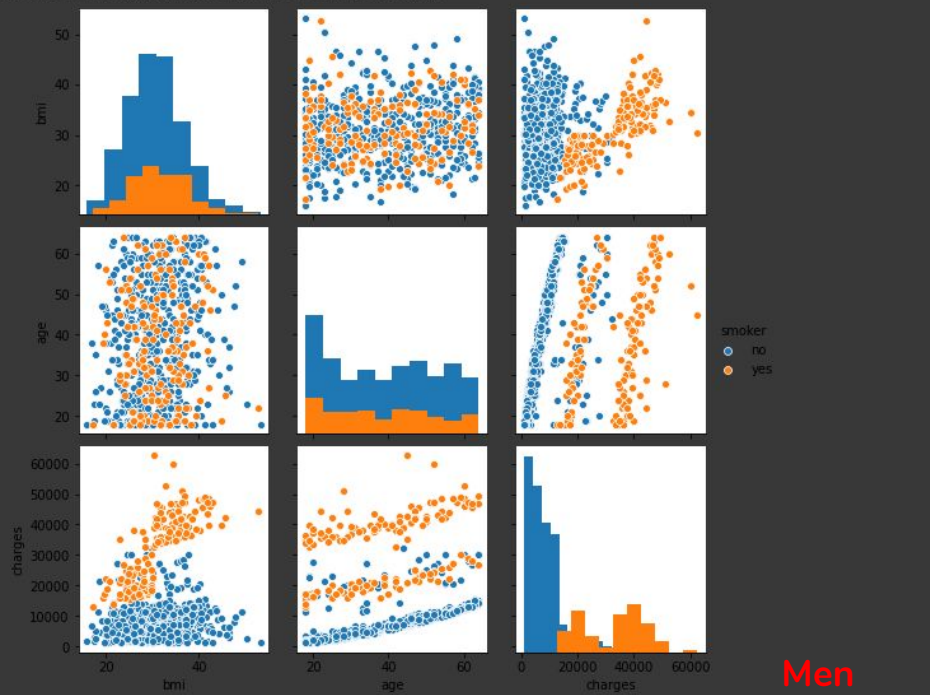
```
df.describe()
```

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010



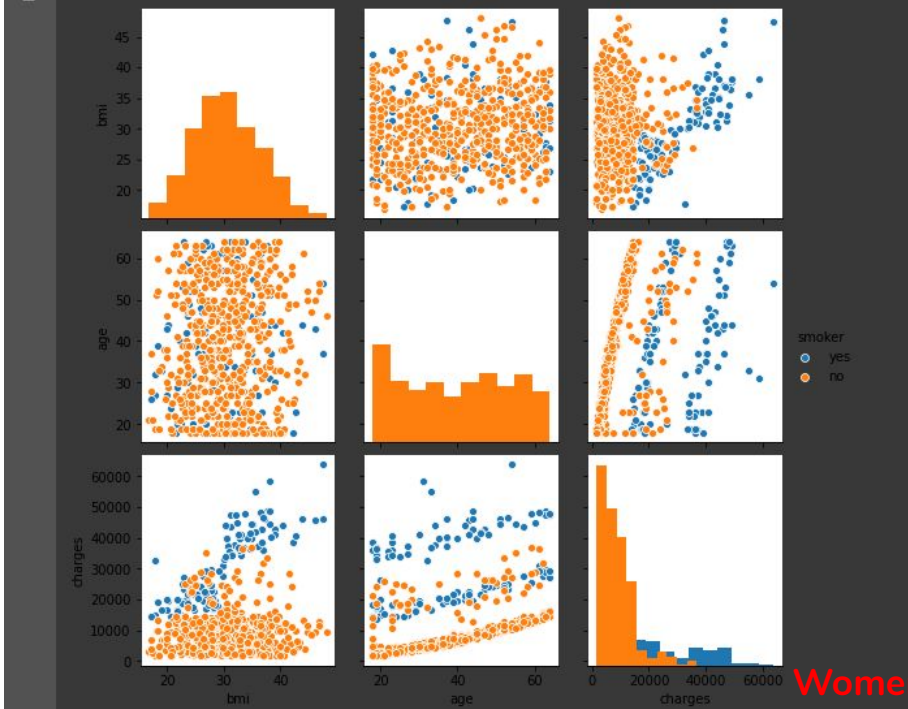
Stage Three

```
sns.pairplot(men, vars=['bmi', 'age', 'charges'], hue="smoker", diag_kind='hist')  
  
ERROR! Session/line number was not unique in database. History logging moved to new session 68  
<seaborn.axisgrid.PairGrid at 0x7f74d50a3400>
```



Men

```
sns.pairplot(women, vars=['bmi', 'age', 'charges'], hue="smoker", diag_kind='hist')  
  
<seaborn.axisgrid.PairGrid at 0x7f74d1f3c860>
```

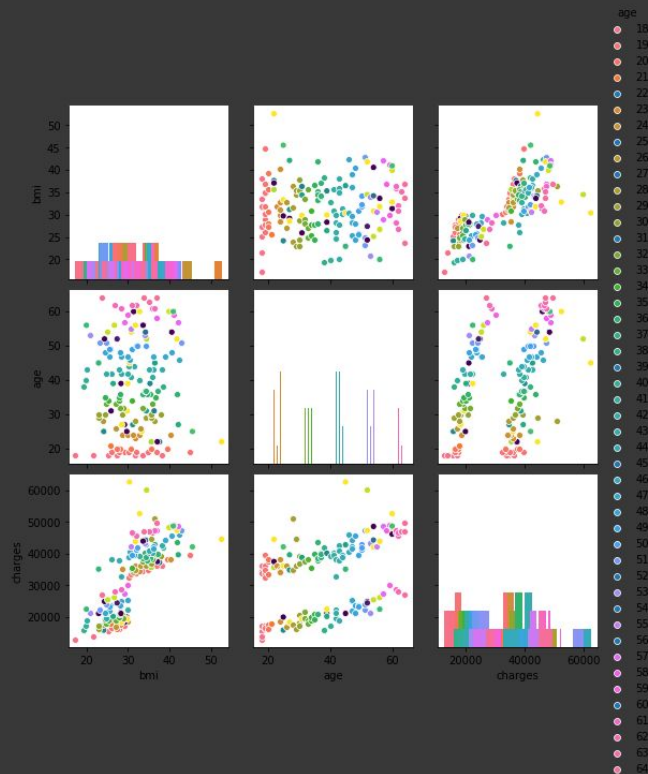


Women

Stage Three

```
[32] sns.pairplot(men_smoker, vars=['bmi', 'age', 'charges'], hue="age", diag_kind='hist')
```

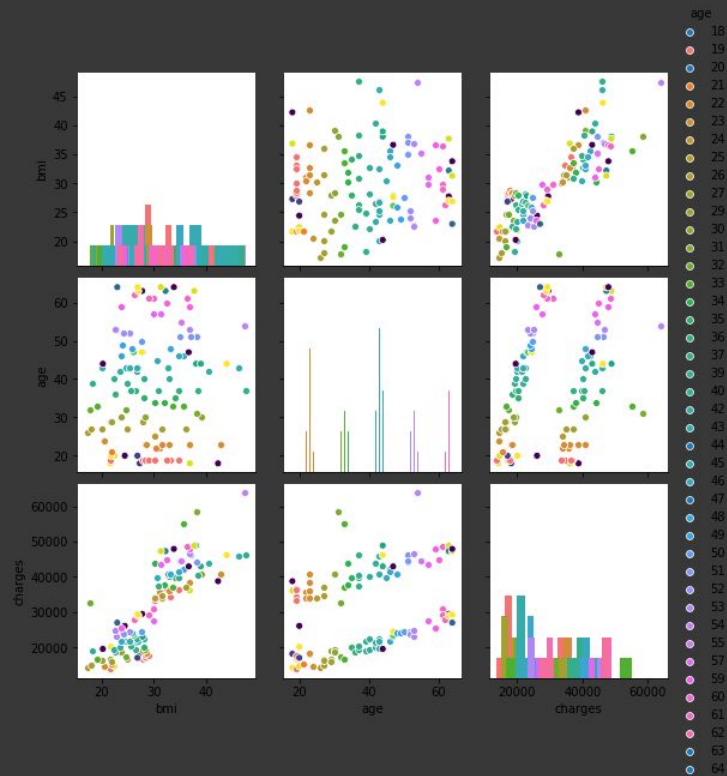
```
<seaborn.axisgrid.PairGrid at 0x7f74d0473b70>
```



Men

```
[31] sns.pairplot(women_smoker, vars=['bmi', 'age', 'charges'], hue="age", diag_kind='hist')
```

```
<seaborn.axisgrid.PairGrid at 0x7f74d0353e10>
```



Women