Axis Insurance

Exploratory Data and Statistical Analysis

PG-DSBA Project 2
Eric Green
December 2020

Objectives

- Explore the dataset and extract insights (EDA)
- Prove (or disprove) that the medical claims made by the people who smoke is greater than those who don't?
- Prove (or disprove) with statistical evidence that the BMI of females is different from that of males
- Is the proportion of smokers significantly different across different regions?
- Is the mean BMI of women with no children, one child, and two children the same?

Data Summary

Data columns (total 7 columns):

- 0 age 1338 non-null int64
- 1 sex 1338 non-null category
- 2 bmi 1338 non-null float64
- 3 children 1338 non-null category
- 4 smoker 1338 non-null category
- 5 region 1338 non-null category
- 6 charges 1338 non-null float64

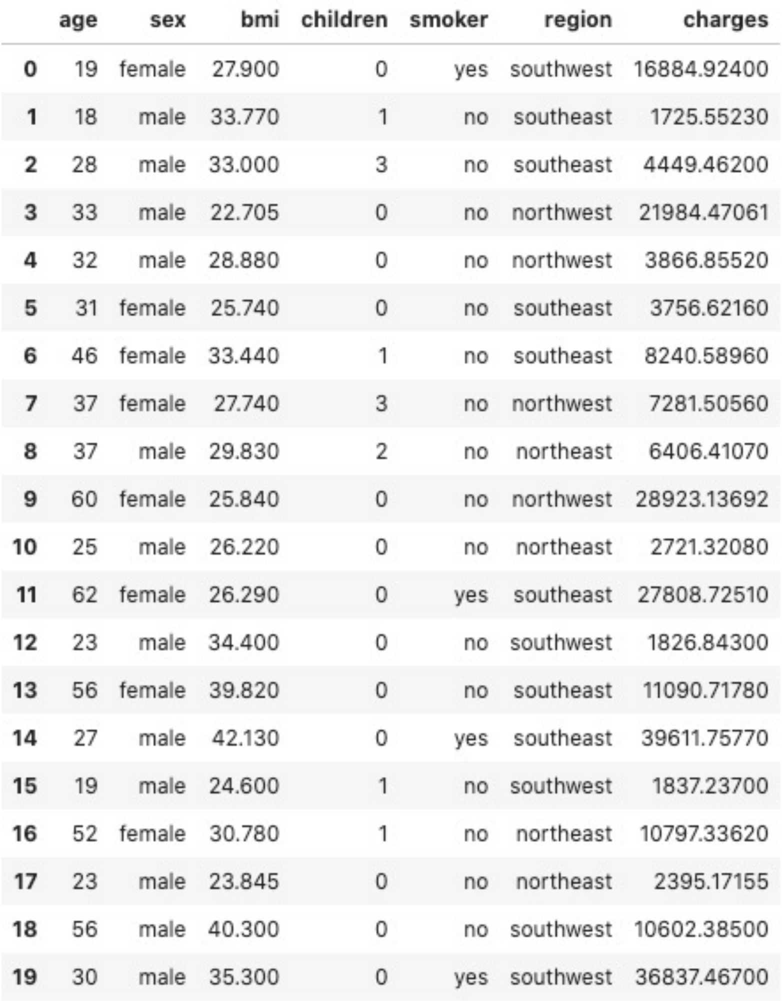
memory usage: 37.3 KB

Data is tidy and clean in raw form. String objects converted to categories to save space.

Total rows: 1338

- Males: 676
- Females: 662
- Smokers: 274
- Nonsmokers: 1064
- Region - northeast: 324
- Region - northwest: 325
- Region - southeast: 364
- Region - southwest: 325

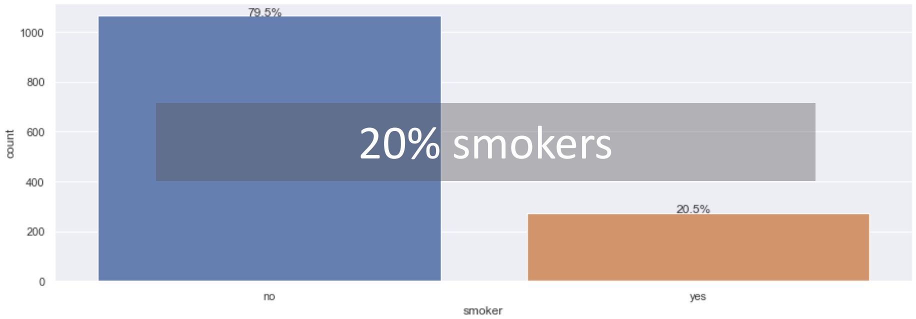
Tidy Data (insurance.csv)



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692
10	25	male	26.220	0	no	northeast	2721.32080
11	62	female	26.290	0	yes	southeast	27808.72510
12	23	male	34.400	0	no	southwest	1826.84300
13	56	female	39.820	0	no	southeast	11090.71780
14	27	male	42.130	0	yes	southeast	39611.75770
15	19	male	24.600	1	no	southwest	1837.23700
16	52	female	30.780	1	no	northeast	10797.33620
17	23	male	23.845	0	no	northeast	2395.17155
18	56	male	40.300	0	no	southwest	10602.38500
19	30	male	35.300	0	yes	southwest	36837.46700

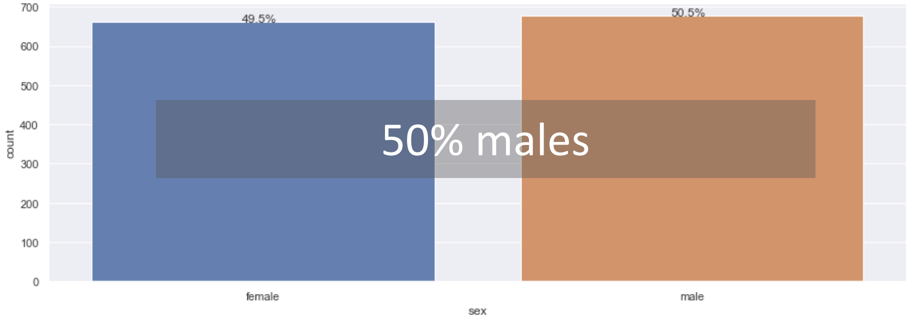
Basic Proportions

Smokers



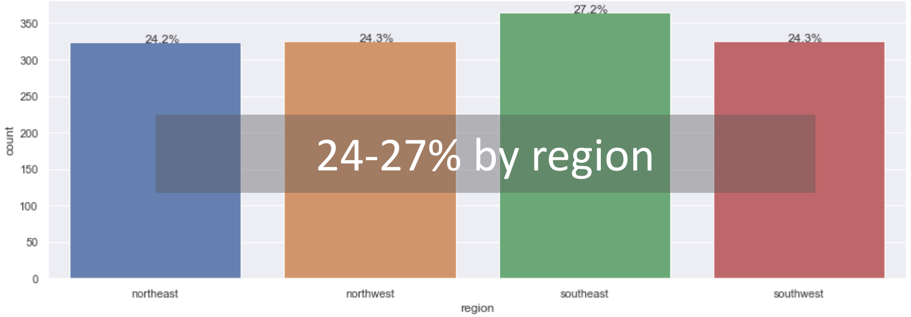
20% smokers

Sex



50% males

Regions



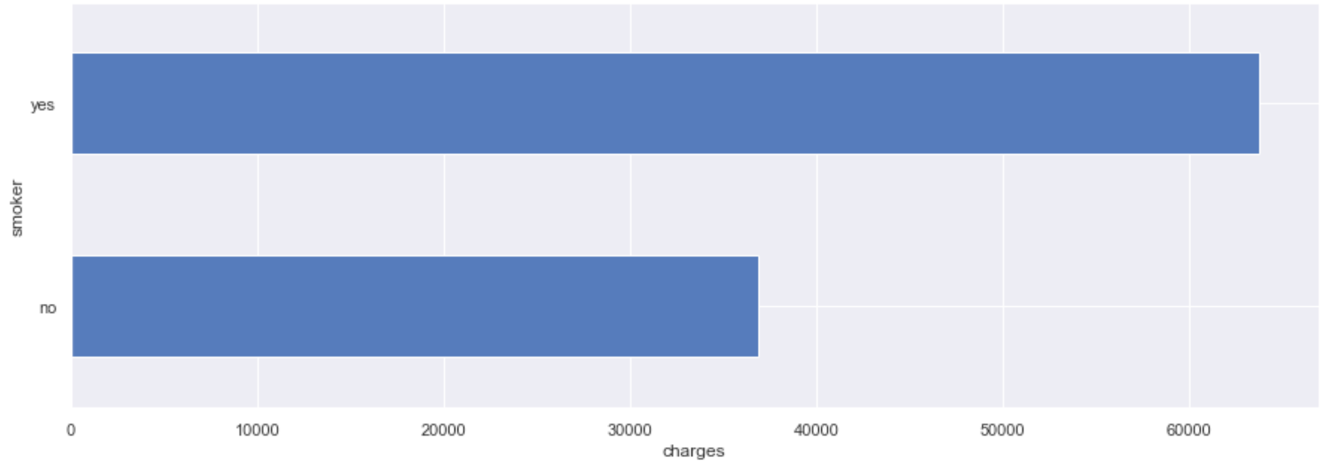
24-27% by region

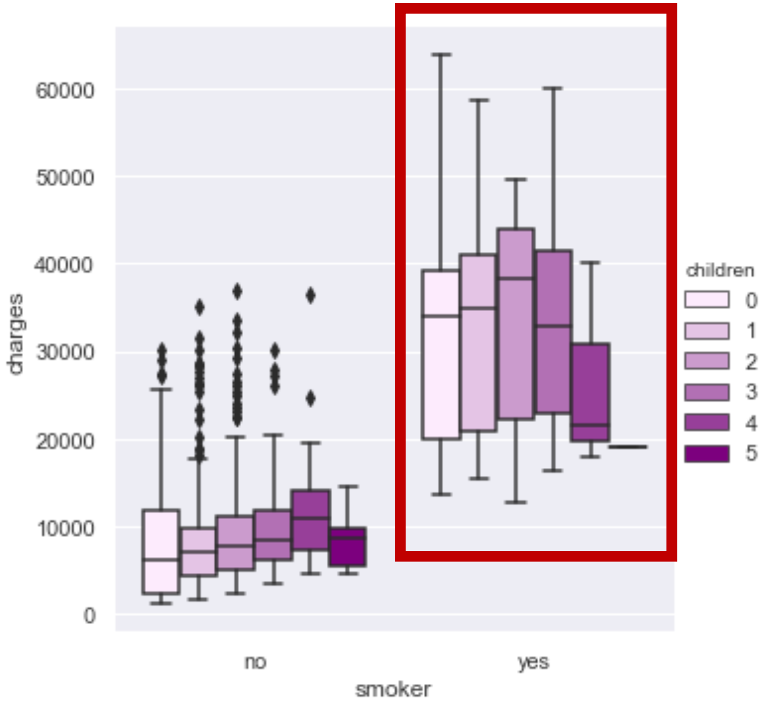
#1 – Hypothesis Testing

Medical charges are higher for smokers vs nonsmokers

Observation

- Smokers' claims/charges are clearly higher than nonsmokers regardless of number of children



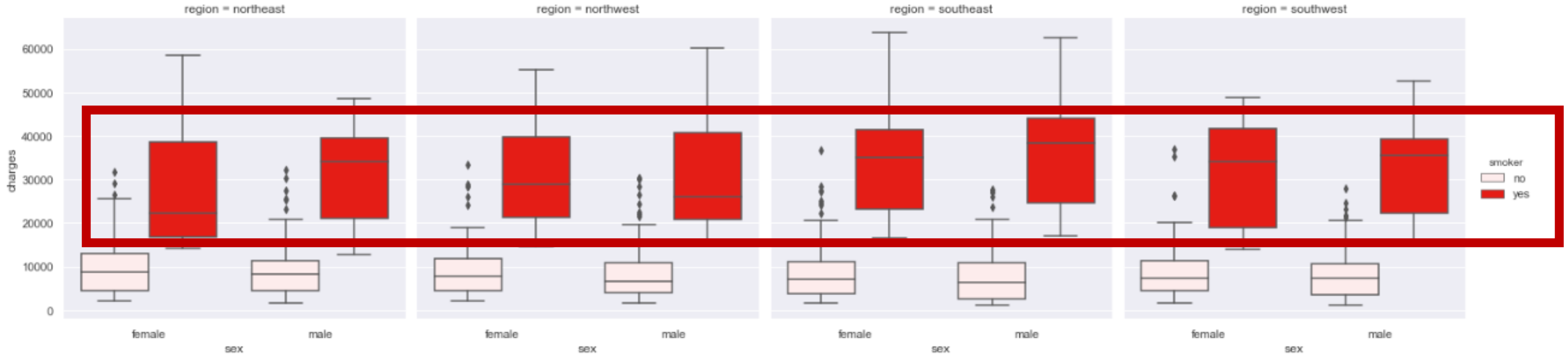


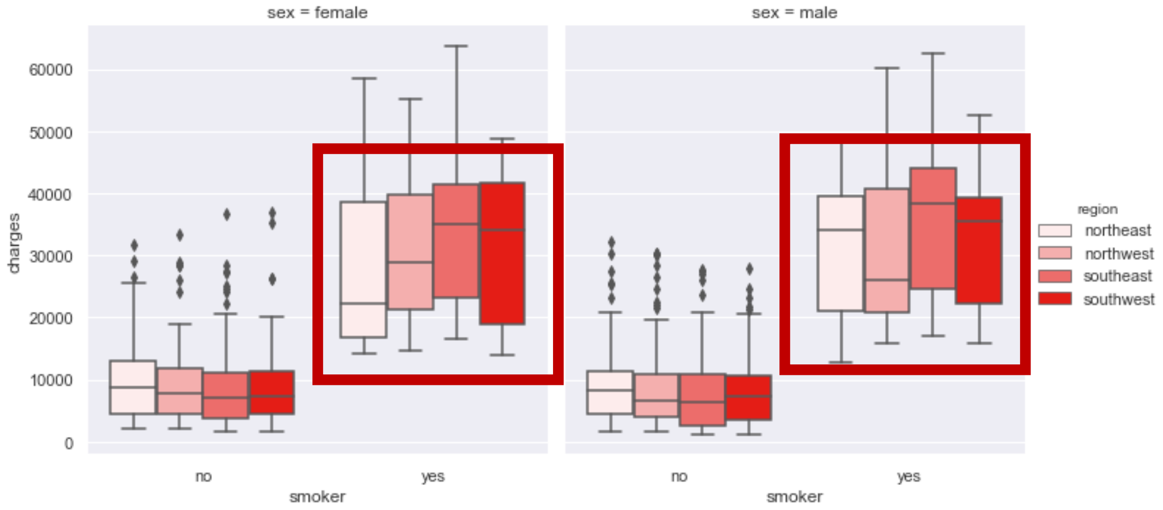
#1 – Hypothesis Testing

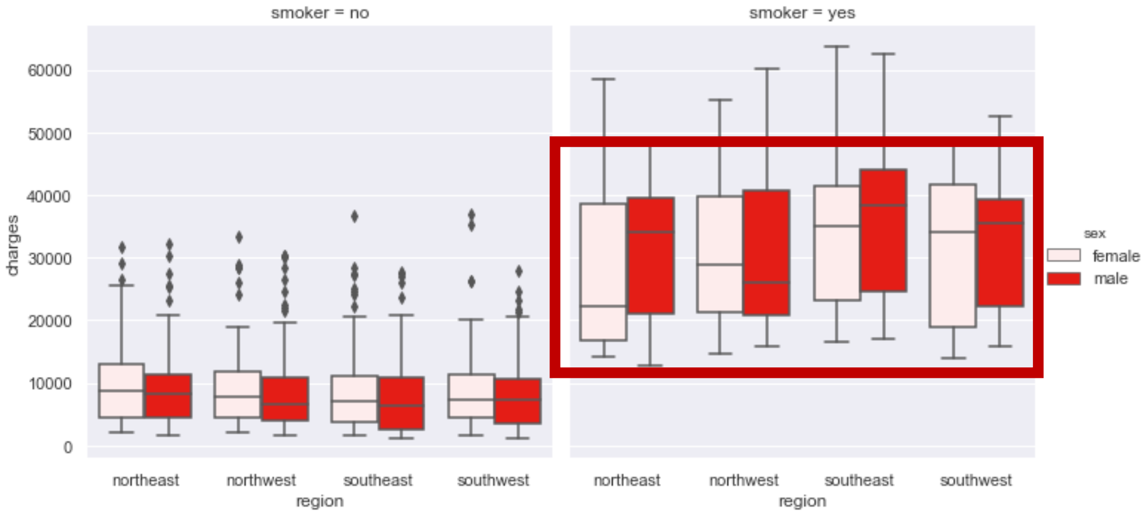
Medical charges are higher for smokers vs nonsmokers

Observation

- Smokers' claims/charges are clearly higher than nonsmokers regardless of sex or region







#1 – Hypothesis Testing

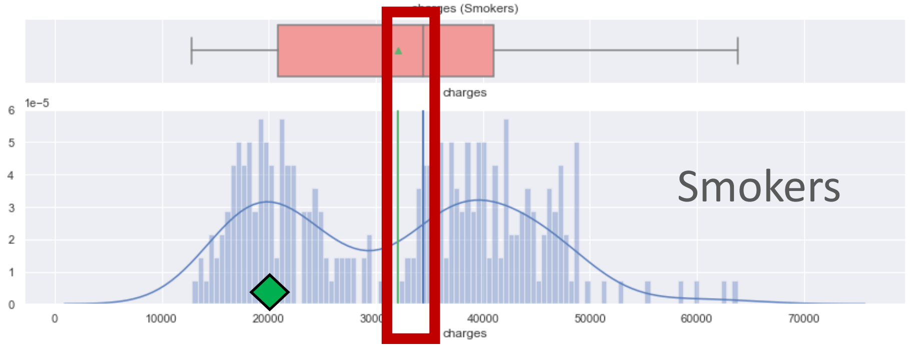
Medical charges are higher for smokers vs nonsmokers

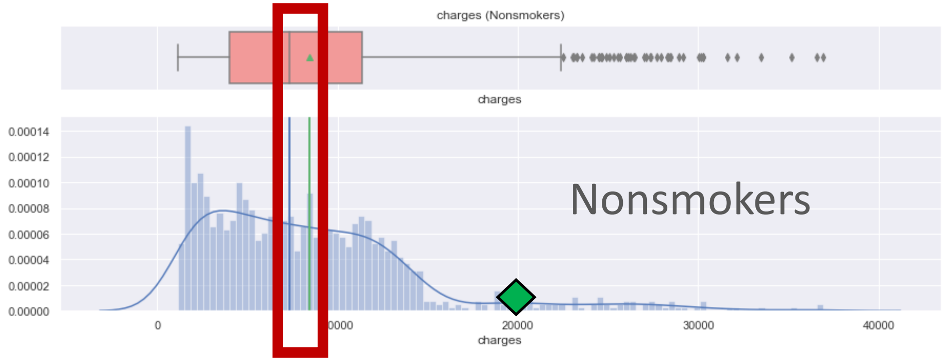
```
*** Basic stats *** Charges by Smoker/Nonsmoker
smoker charges:
count 274.000000
mean 32050.231832
std 11541.547176
min 12829.455100
25% 20826.244213
50% 34456.348450
75% 41019.207275
max 63770.428010

Name: charges, dtype: float64
nonsmoker charges:
count 1064.000000
mean 8434.268298
std 5993.781819
min 1121.873900
25% 3986.438700
50% 7345.405300
75% 11362.887050
max 36910.608030
Name: charges, dtype: float64

*** T-TEST *** charges by smokers
t-statistic result: 46.664921
p-value result: 8.271435842177219e-283
```

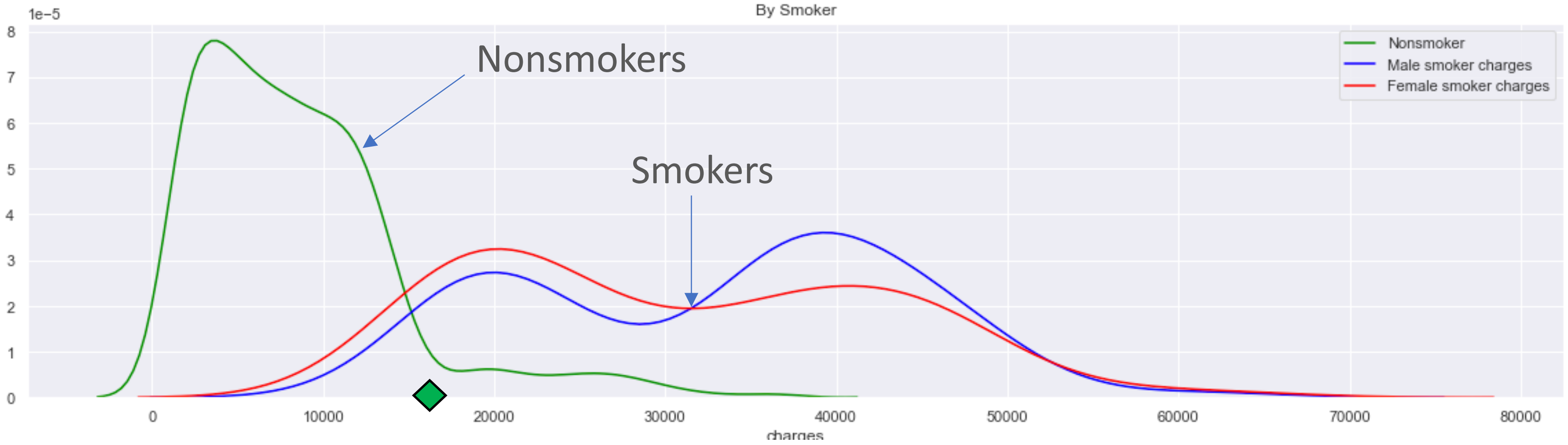
P-value indicates means and variances (of charges) of smokers vs nonsmokers are very different. Delta of 46.7 SDs, e-284 probability

Observations

- Clearly, smokers incur much higher medical claims/charges than do nonsmokers, regardless of any other variable present
- Distribution means and medians are visually shifted/offset
- We can accept the hypothesis that smoker’s medical charges are more than nonsmokers
- While visually apparent, the statistical tests show these distributions share very little in common and are therefore not the same (sameness is rejected)
- Interesting here that smokers represent only 20% of the full sample







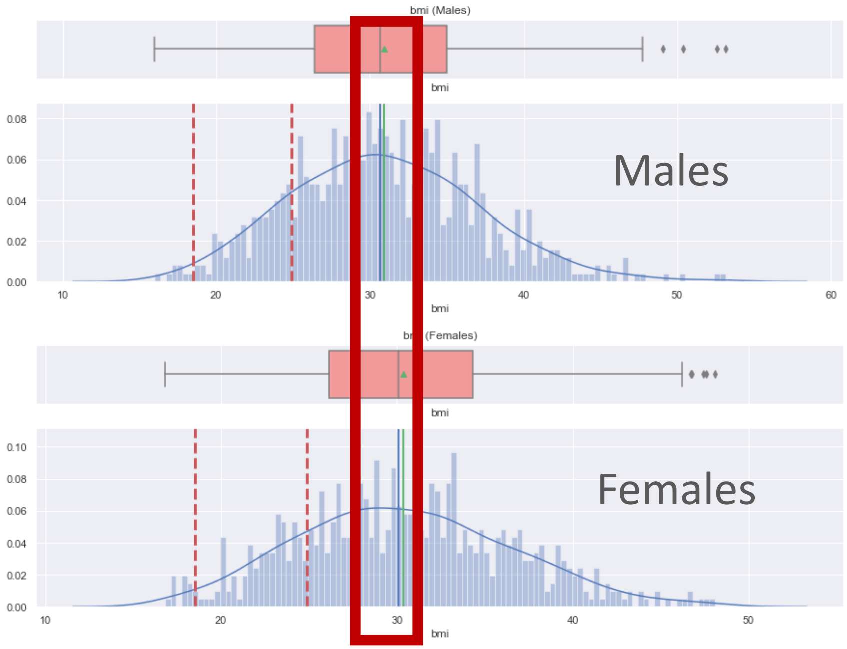
#2 – Hypothesis Testing

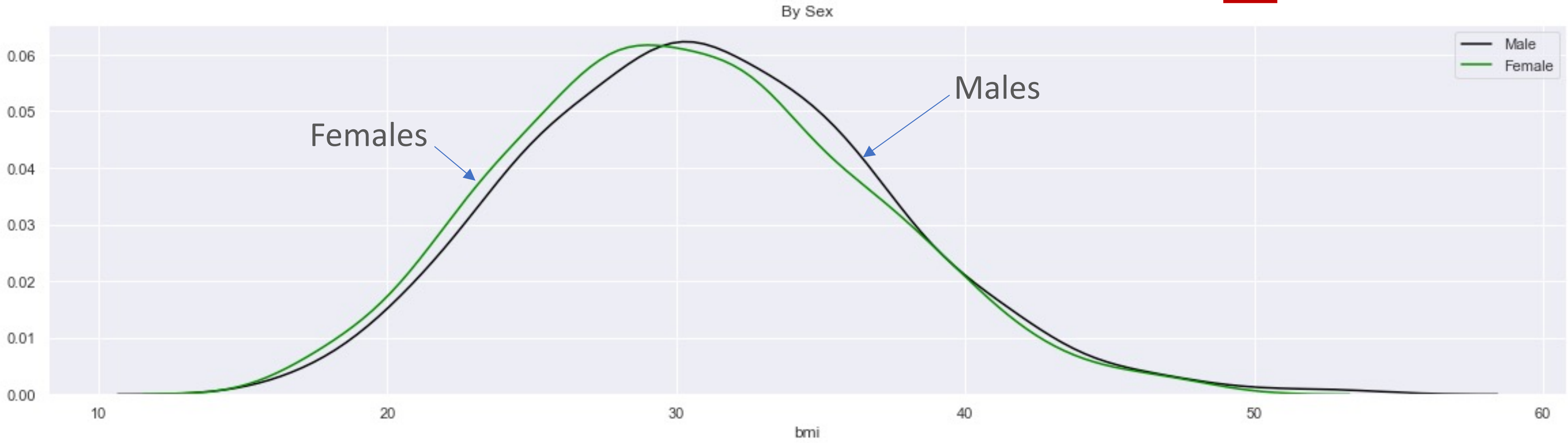
BMI of females is different than males

Observations

- Visually, the 2 distributions look pretty similar in their shape and their respective means and medians being close together
- The KDE plot overlays the silhouette traces which shows the curves to be normal and closely aligned, visually
- A statistical T-test indicates similarity between the males and female distributions (1.7 SDs mean delta, 8.9% probability of similarity)
- Based on the evidence, we conclude that BMI for males and females is the same (sameness accepted)

```
*** T-TEST *** bmi by sex
t-statistic: 1.696753
p-value: 0.08997637178984932
```





#3 – Hypothesis Testing

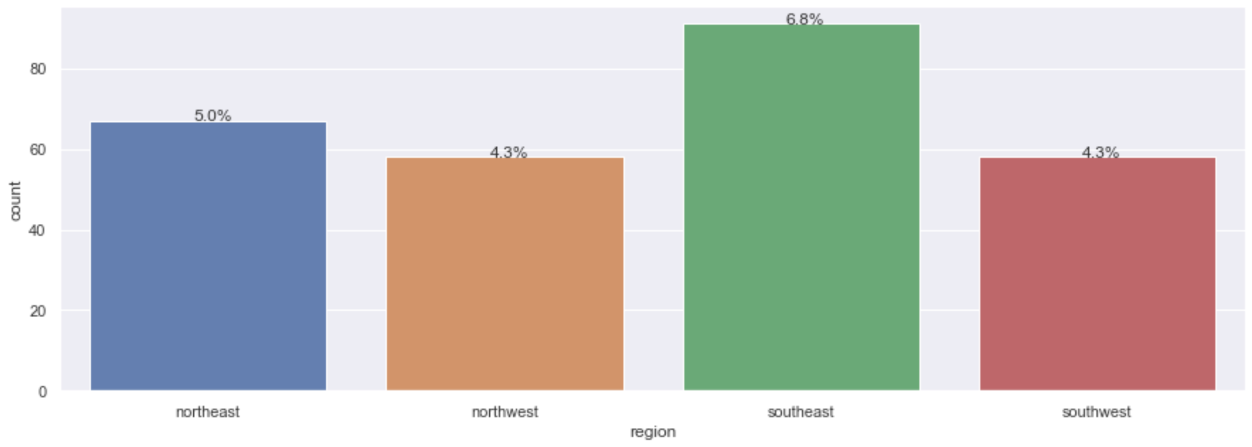
Proportions of smokers is different across regions

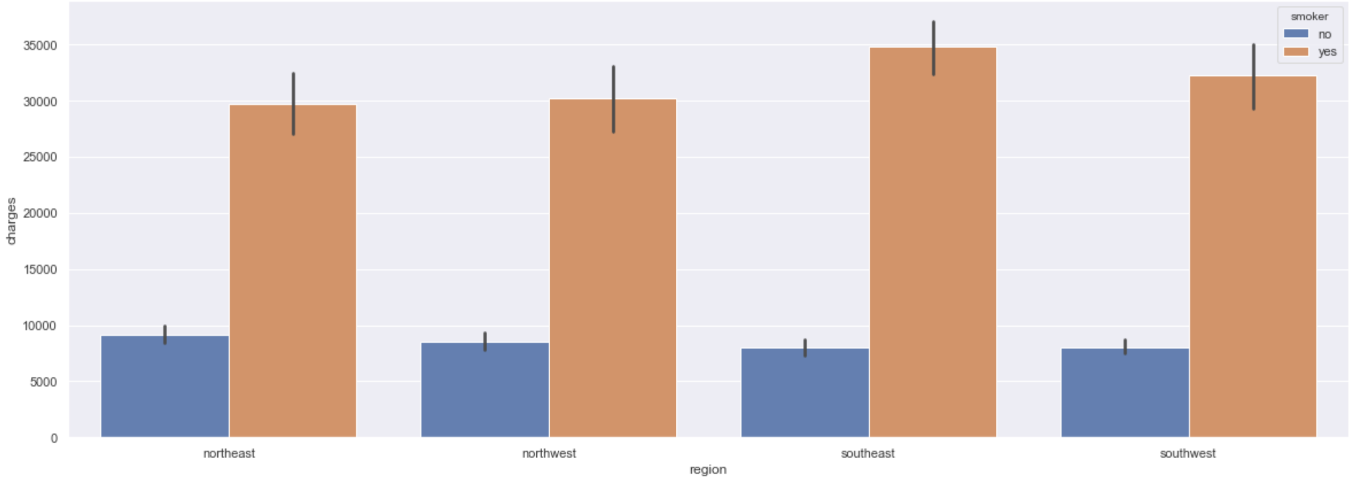
Observations

- Visually, the proportions of smokers across the regions has some variation and also some similarities
- Majority of charges by region made by smokers
- It depends what we want to understand to select a statistical test – each Test
  - Compares frequencies against equally proportioned frequencies
  - GOF test indicates similarity of .013, which crosses our significance level of .05 (different, not equally proportioned)
  - Chi2 test indicates a p-value of .176, which stays left of our significance level .05 (same, equally proportioned)
  - The test results conflict because they each land on opposite sides of the significance level
  - We should refine the line of questioning here and do further analysis

```
*** GoF TEST *** Smokers by Region
Observed: 67 58 91 58
Expected: 68.5 68.5 68.5 68.5
chi squared: 10.642336
p-value: 0.01382579480288941

*** chi2_contingency TEST *** Smokers by Region
obs: [[67. 58. 91. 58. ] [68.5 68.5 68.5 68.5]]
(4.9336693612268, 0.17671913436450915, 3,
  array([[67.75, 63.25, 79.75, 63.25], [67.75, 63.25, 79.75,
63.25]]))
```

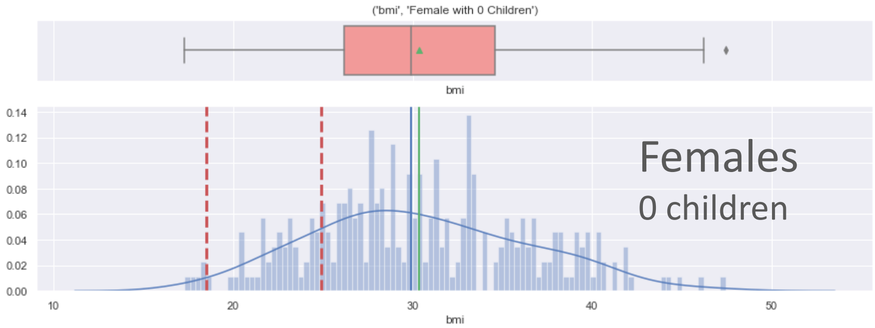
Smokers by Region





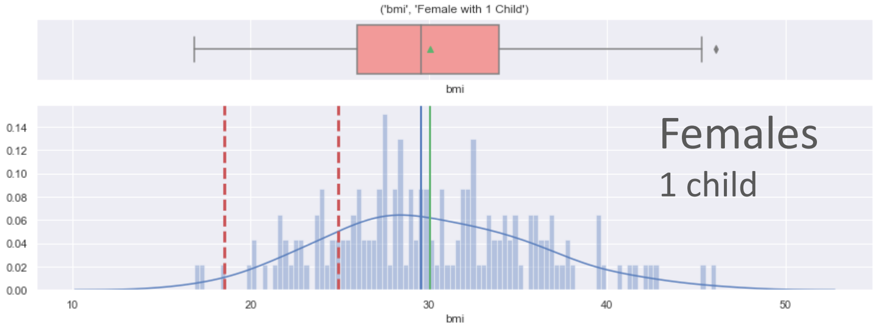
#4 – Hypothesis Testing

Female mean BMI by children (0, 1, 2)

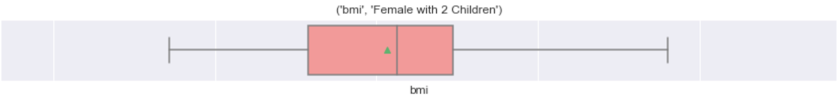


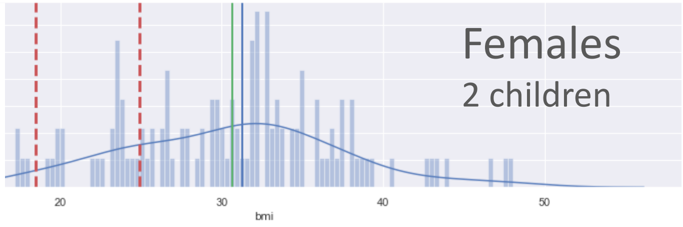
Observations

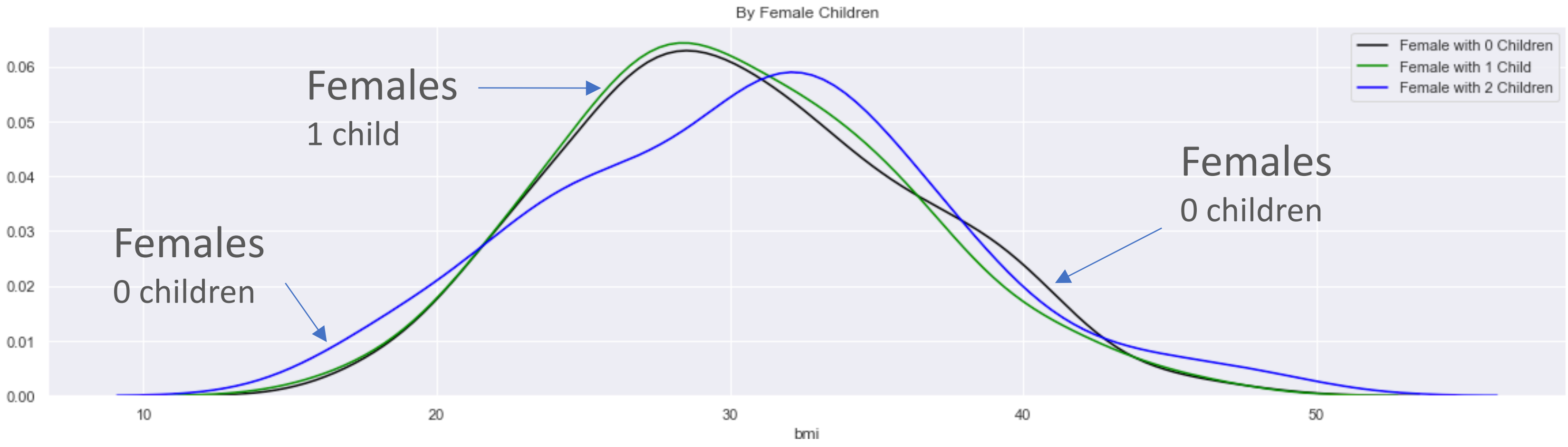
- Visually, the proportions of smokers across the regions has some variation and also some similarities in shape, central tendency and dispersion
- The 3 curves appear to be normally distributed and somewhat aligned, visually
- The ANOVA test indicates .71 similarity between the actual proportions and equal proportions
- We can conclude that BMI mean and variance for the 3 groups are similar, regardless of # of children (sameness accepted)



```
*** ANOVA TEST *** Female BMI mean by children 0,1,2
                sum_sq    df      F          PR(>F)
C(Kids) 24.590123 2.0 0.334472 0.715858
Residual 20695.661583 563.0 NaN NaN
```







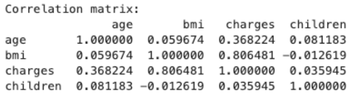
Means

children 0: 30.361522491349486
children 1: 30.05265822784811
children 2: 30.64978991596637

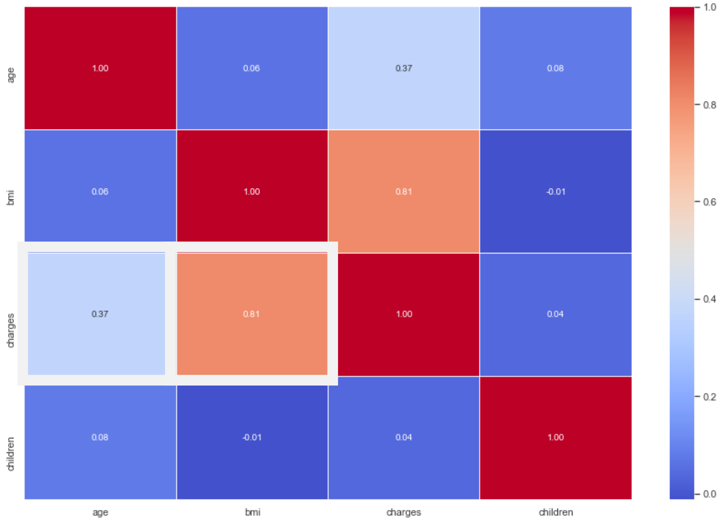
Risk Variable – BMI / Charges

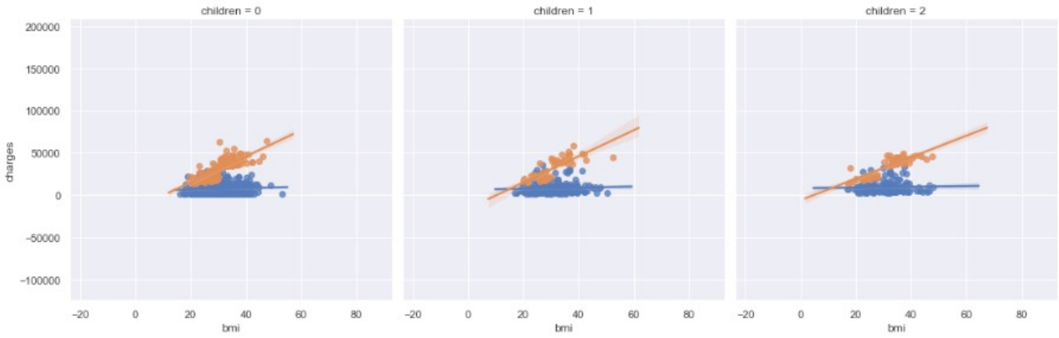
Observation

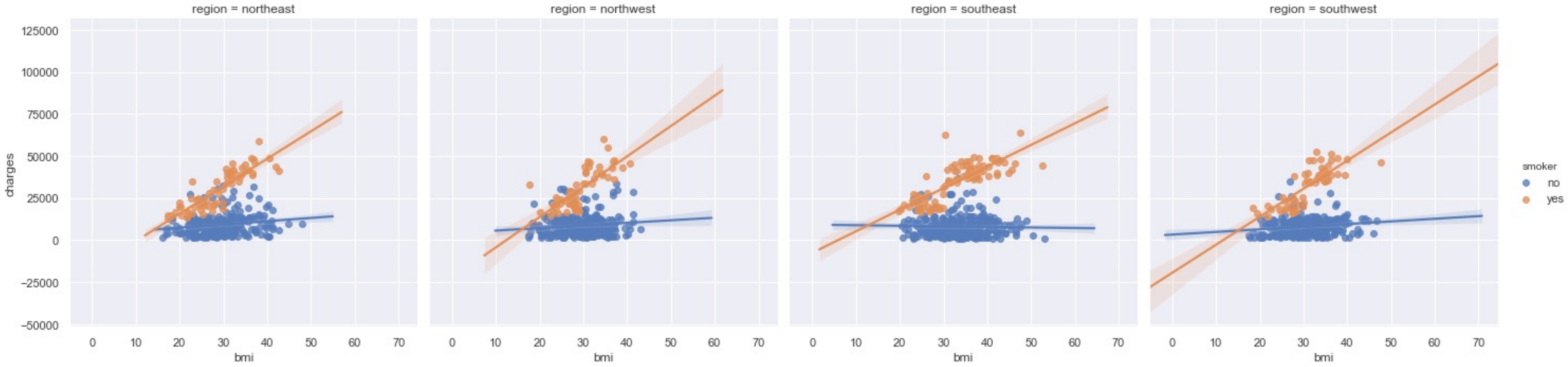
- Strong positive correlation between BMI and medical charges can be seen in both linear plot and correlation heatmap
- Weak correlation between age and charges
- Idea: model/optimize a risk score for future ML uses



```
Correlation matrix:
               age       bmi   charges  children
age       1.000000  0.059674  0.368224  0.081183
bmi       0.059674  1.000000  0.806481 -0.012619
charges   0.368224  0.806481  1.000000  0.035945
children  0.081183 -0.012619  0.035945  1.000000
```







Conclusions

- Smokers incur the majority of claims/charges vs. nonsmokers
- 20% of the sample account for majority of claims/charges
- BMI for males and females is similar (p-value.09>.05)
- Proportions of smokers across regions have approximate similarities. Depending on which statistical test is used, Ho (null) or H1 (alt) can be accepted at .05 significance
    - Inconclusive - adjust significance level based on more questioning
- BMI for females across 0, 1 or 2 children is highly similar and not statistically different in terms of mean & variance (sameness accepted)
- BMI is strongly correlated with charges across all groups
- BMI trends can be used in forecasting claims/charges

Recommendations to Business

Tactical

1. Address line of questioning regarding smoker proportions across regions (try alpha .01)
2. Investigate pricing optimization using BMI data
3. Investigate “insurability” criteria for future applicants
4. Perform risk analysis on current data
5. Investigate increasing BMI/charges data sampling

Strategic

Future R&D

- Risk audit - profile ratios in current data of bmi/charges
- Price tuning & optimization
    - (scale pricing with BMI & claims/charges)
- Risk Score development & screening optimization
- Claims forecasting engine