COGS 108 Group project
Analysis of possible relationship between medical expenditure in the US and potential factors

Group 23: Yuhan Zhou, Baoni Li, Huiyi He,
Jiayi Zhu and Yihuan Wang

Overview, Question & Background

- “The United States has one of the highest costs of healthcare in the world. “

- Healthcare expenditure is a critical indicator of a country’s health status. There are many features are associated with higher medical expenditure.

- If people can predict their body changes in the near future based on their body indicators and lifestyles, it will be much easier to select the most appropriate healthcare insurance plans for themselves.

- Using a multivariate OLS model, we found that older people and higher BMI could cause high medical expenditure. Working Hours and smoking behavior have a negative correlation with medical expenditure. Also, an individual’s region and racial characteristics are correlated with medical expenditure.

Research Question

What factors are associated with higher medical expenditure in adults in the United States?

Specifically, are there relationships between medical costs and demographic factors such as age, working hours, and BMI, as well as health-related behaviors such as smoking behavior and overall lifestyle choices?

If there are statistically significant relationships between these features and medical expenditure, are they positively or negatively associated?

Hypothesis

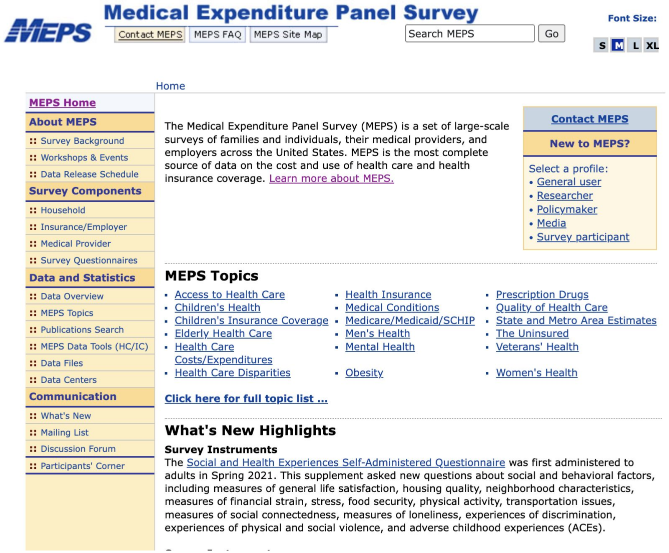
- **Age**
- **Working hours**
- **BMI**
- **Lifestyle differences**
- **Region and racial characteristics**

Data description

Medical Expenditure Panel Survey

National survey conducted by the Agency for Healthcare Research and Quality (AHRQ) that collects data on healthcare utilization and expenditures.

2011-2020



The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. MEPS is the most complete source of data on the cost and use of health care and health insurance coverage. Learn more about MEPS.

MEPS Topics

- Access to Health Care
- Children's Health
- Children's Insurance Coverage
- Elderly Health Care
- Health Care Costs/Expenditures
- Health Care Disparities
- Health Insurance
- Medical Conditions
- Medicare/Medicaid/SCHIP
- Men's Health
- Mental Health
- Obesity
- Prescription Drugs
- Quality of Health Care
- State and Metro Area Estimates
- The Uninsured
- Veterans' Health
- Women's Health

Click here for full topic list ...

What's New Highlights

Survey Instruments

The Social and Health Experiences Self-Administered Questionnaire was first administered to adults in Spring 2021. This supplement asked new questions about social and behavioral factors, including measures of general life satisfaction, housing quality, neighborhood characteristics, measures of financial strain, stress, food security, physical activity, transportation issues, measures of social connectedness, measures of loneliness, experiences of discrimination, experiences of physical and social violence, and adverse childhood experiences (ACEs).

Data cleaning/processing

100,000 observations and 1000 columns —> less than 20 variables

Dummy variables: region, smoking habits, race

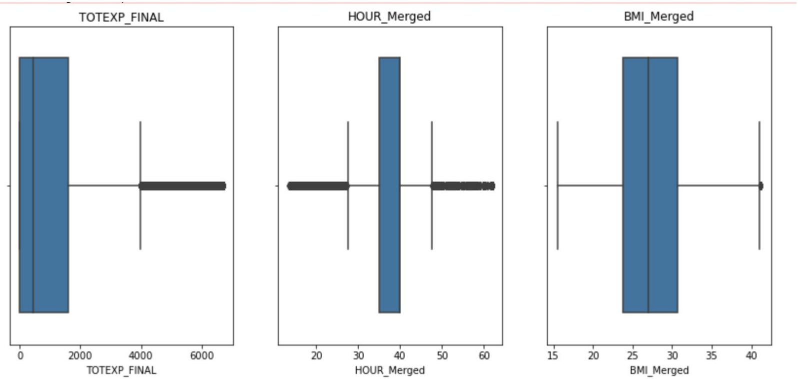
Numerical variables: age, hours worked, medical expenditure, Body Mass Index

Remove Outliers:

total annual medical expenditure(TOTEXP_FINAL)

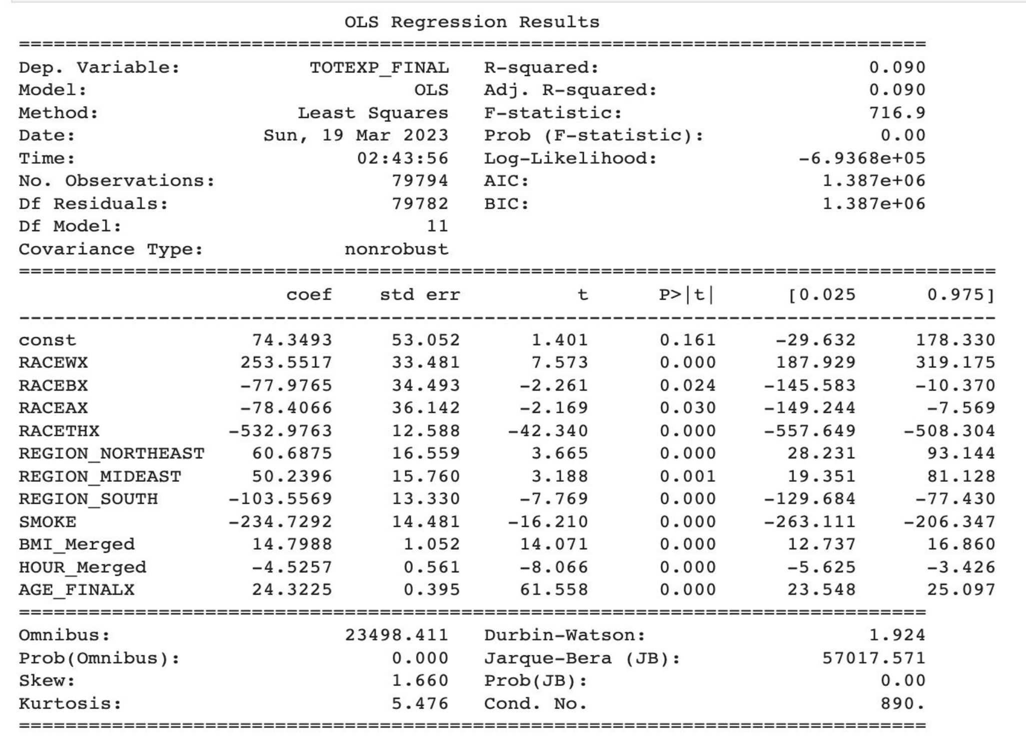
hours worked per week(HOUR_Merged)

Body Mass Index(BMI_Merged)



Data Visualization Analysis and Results

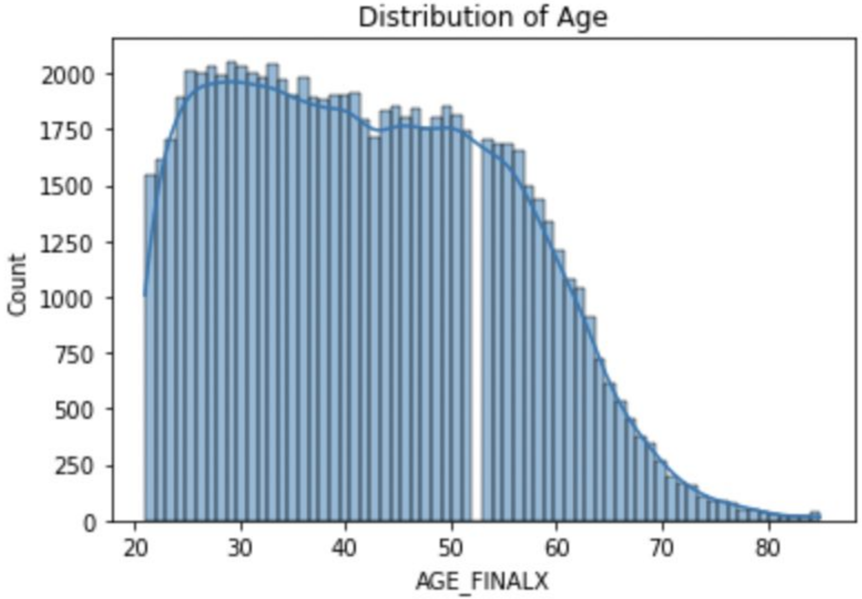
Regression Result



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           TOTEXP_FINAL   R-squared:                       0.090
Model:                            OLS   Adj. R-squared:                  0.090
Method:                 Least Squares   F-statistic:                     716.9
Date:                Sun, 19 Mar 2023   Prob (F-statistic):               0.00
Time:                        02:43:56   Log-Likelihood:            -6.9368e+05
No. Observations:               79794   AIC:                         1.387e+06
Df Residuals:                   79782   BIC:                         1.387e+06
Df Model:                          11
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const               74.3493     53.052      1.401      0.161     -29.632     178.330
RACEWX             253.5517     33.481      7.573      0.000     187.929     319.175
RACEBX             -77.9765     34.493     -2.261      0.024    -145.583     -10.370
RACEAX             -78.4066     36.142     -2.169      0.030    -149.244      -7.569
RACETHX           -532.9763     12.588    -42.340      0.000    -557.649    -508.304
REGION_NORTHEAST    60.6875     16.559      3.665      0.000      28.231      93.144
REGION_MIDEAST      50.2396     15.760      3.188      0.001      19.351      81.128
REGION_SOUTH      -103.5569     13.330     -7.769      0.000    -129.684     -77.430
SMOKE             -234.7292     14.481    -16.210      0.000    -263.111    -206.347
BMI_Merged          14.7988      1.052     14.071      0.000      12.737      16.860
HOUR_Merged         -4.5257      0.561     -8.066      0.000      -5.625      -3.426
AGE_FINALX          24.3225      0.395     61.558      0.000      23.548      25.097
==============================================================================
Omnibus:                    23498.411   Durbin-Watson:                   1.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            57017.571
Skew:                           1.660   Prob(JB):                         0.00
Kurtosis:                       5.476   Cond. No.                         890.
==============================================================================
```

Age and Medical Expenditure

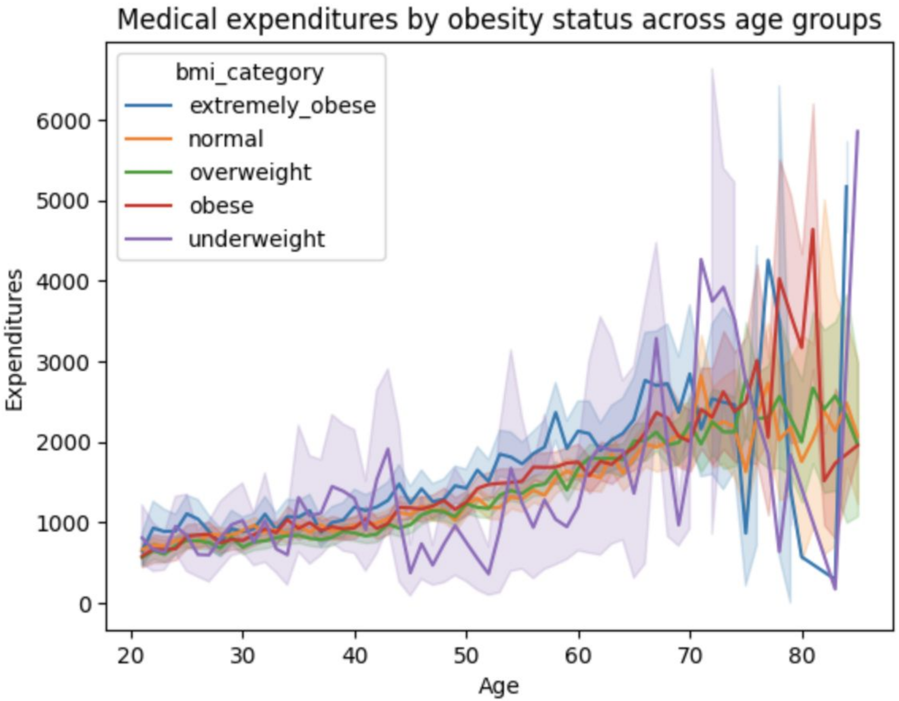
- AGE is statistically significant with value of 24.3225, implies keeping all else constant.
- Our model estimates on average a 24.3225$ increase on medical expenditure with one additional year of age



BMI and Medical expenditures

- We concluded that people with more extreme BMI values tend to have higher medical expenditures across all age groups.

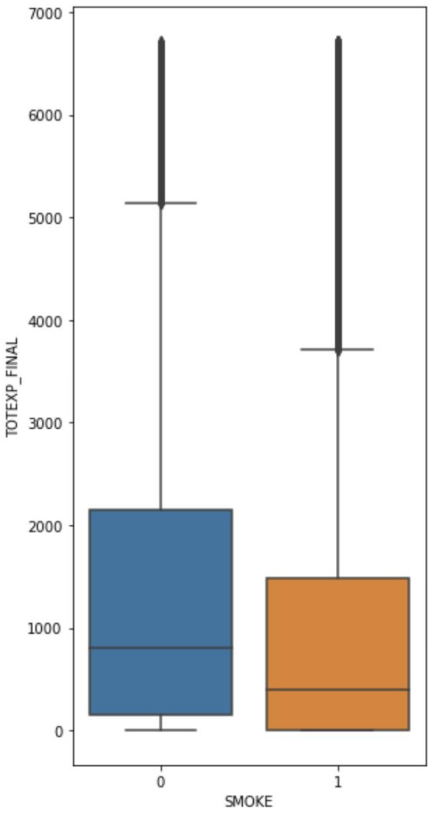
- From the regression results, we observe that there is a positive relationship between BMI and medical expenditure. On average, people with higher BMI values have higher medical expenditures



Smoke and Medical expenditure

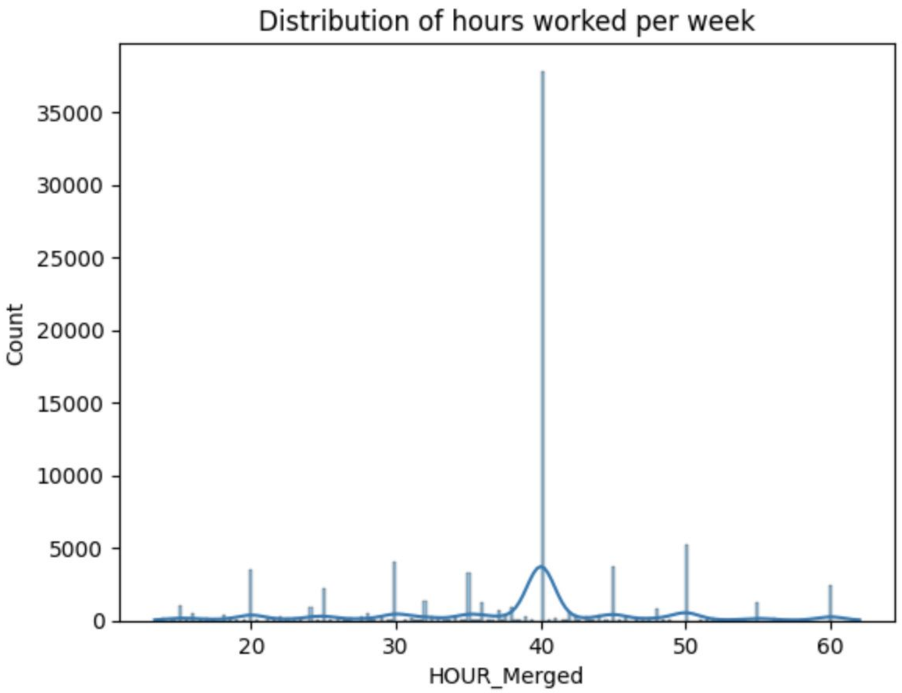
- The mean expenditure of people not smoke is about 800, but that of people smoke is about 500. In general, people not smoke will cost more expenditure.

- In our OLS regression result, we can see that the coefficient of smoke feature is -234.7292, implying that smoking has a negative relationship with total expenditure.



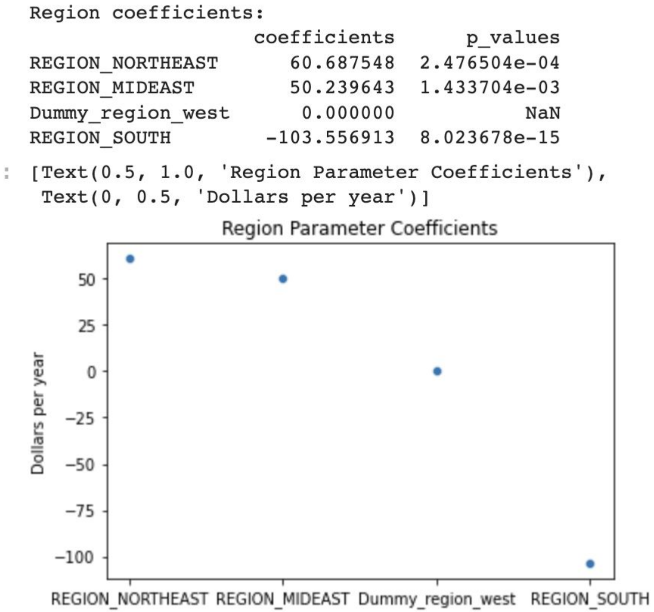
Hours worked per week and Medical expenditure

- We observe that people's working hours tend to concentrate on 20, 30, 40, 50 and 60 hours
- We conclude that there is a negative correlation between working hours and medical expenditures.



Region and Medical expenditure

- On average, individuals live in the Northeast region of the nation have the highest medical spending - 61 dollar more per year than those who live in the west, where as those who lives in south have the lowest average medical spending - 104 less than those who live in the west.

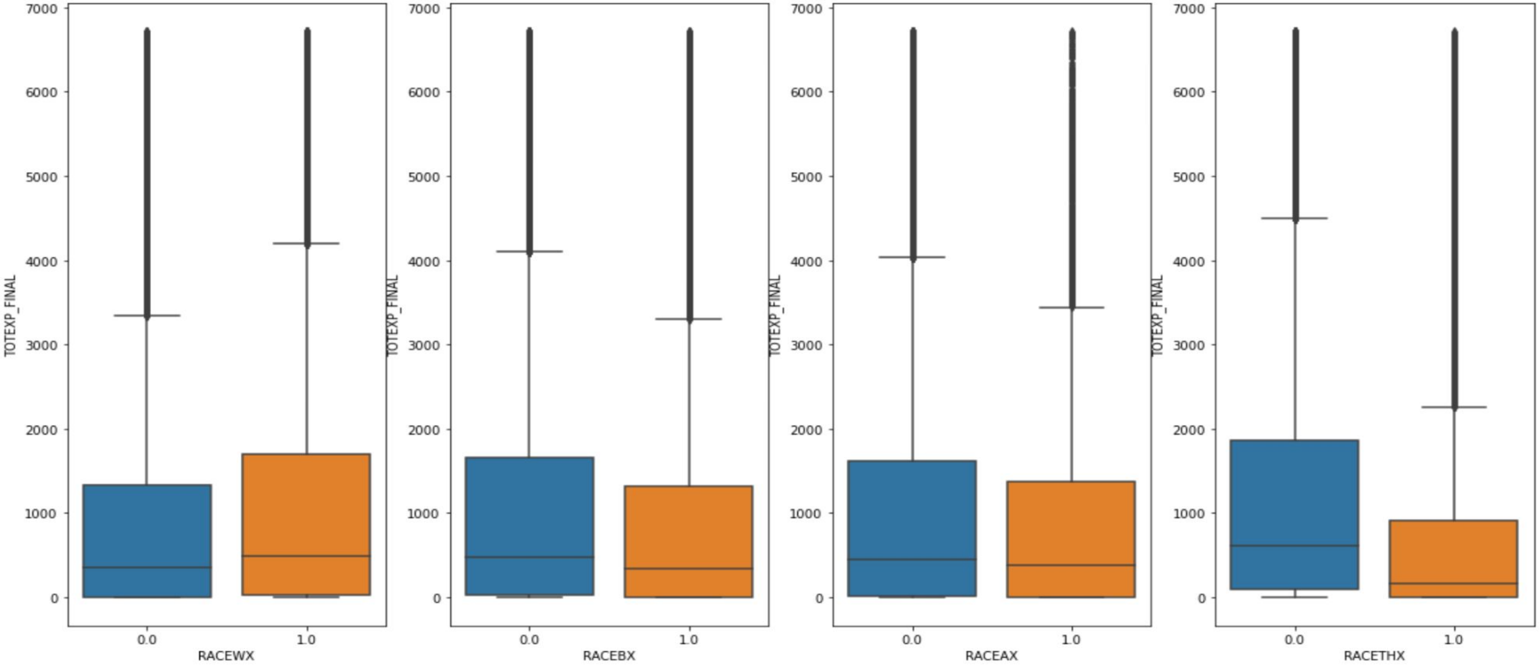


```
Region coefficients:
                      coefficients      p_values
REGION_NORTHEAST         60.687548  2.476504e-04
REGION_MIDEAST           50.239643  1.433704e-03
Dummy_region_west         0.000000           NaN
REGION_SOUTH           -103.556913  8.023678e-15
```

```
: [Text(0.5, 1.0, 'Region Parameter Coefficients'),
   Text(0, 0.5, 'Dollars per year')]
```

Dot plot showing estimator coefficients for all dummy region variables in ascending order

Race and Medical expenditure



Distribution of medical expenditure by race

Ethics and Privacy

- Ethical concerns regarding the dataset and bias in the data

- Analysis, post-analysis and communication

Conclusion and Discussion

Thanks!