Group05_Report

Feifei Li | Izzy Mika

STAT 311 Regression Analysis, Fall 2025

Modeling U.S. Health Insurance Charges:

The Joint Effects of BMI and Smoking

Introduction

Understanding which demographic and lifestyle factors drive U.S. health insurance charges is essential for effective risk assessment and premium setting. Smoking and obesity are known to contribute heavily to medical expenditures, but it remains unclear whether BMI affects smokers and non-smokers differently. This project investigates that question using an observational dataset of 1,338 insured individuals.

Research Questions

1. How do age, BMI, number of children, and smoking status influence annual medical insurance charges?
2. Does BMI affect charges differently for smokers versus non-smokers?
3. Does applying a log transformation to charges improve model adequacy given the strong right-skewed distribution?

Significance

The findings are relevant to:

- Insurers, who rely on accurate pricing models
- Policymakers, who allocate preventive health resources
- Consumers, who want to understand how lifestyle factors influence financial risk

The BMI × smoker interaction is especially important because smoking modifies metabolic and cardiovascular risks related to body weight, potentially altering its financial impact.

Approach

We followed a structured modeling process:

- Data screening and cleaning
- Exploratory Data Analysis (EDA)
- Variable screening (stepwise and all-possible-regressions)
- Model building (three nested models)
- Diagnostic testing
- Model validation (train/test split, PRESS)
- Final model selection

All analyses were conducted in JMP.

Data Description

The Health Insurance Charges dataset from Kaggle contains 1,338 observations and seven variables: age, sex, BMI, children, smoker, region, and annual medical charges. After removing three incomplete records, 1,335 observations remained.

Quantitative variables: age, bmi, children, charges

Qualitative variables: sex, smoker, region

(Smoker is coded such that **"yes" is the baseline** and **smoker = no** is the indicator.)

The dataset is pre-structured, with no evidence of heavy preprocessing, and follows a typical cross-sectional insurance-pricing format in which each record reflects one individual's demographic and lifestyle characteristics and their associated annual insurance cost.

The response variable charges is strongly right-skewed with many legitimate high-cost observations, a pattern common in healthcare expenditure data. The predictors represent factors commonly used in real insurance underwriting, allowing meaningful exploration of how demographic and lifestyle variables influence medical expenses.

Exploratory Data Analysis (EDA)

Quantitative Variable Exploration (See figure 1 in Appendix)

age: Uniform adult distribution, mild right skew
bmi: Right-skewed , range 16-53
c**hildren:** Right-skewed count variable, mostly 0–2.
c**harges:** Extremely right-skewed with a long tail, indicating potential need for log transform.

Distributional Summaries

Variable	Mean	Std. Dev	Min	Max
age	~39	~14	18	64
bmi	~30.7	~6.2	15.96	53.13

Variable	Mean	Std. Dev	Min	Max
children	~1.09	~1.21	0	5
charges	~13,270	~12,110	1,122	63,770

Qualitative Variable Exploration

Sex: Balanced; little difference in charges.

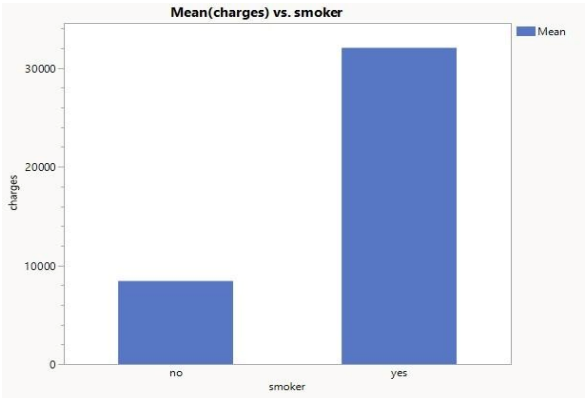
Region: Balanced; small charge differences across regions.

Smoker: Only ~20% smokers, but they show dramatically higher charges.

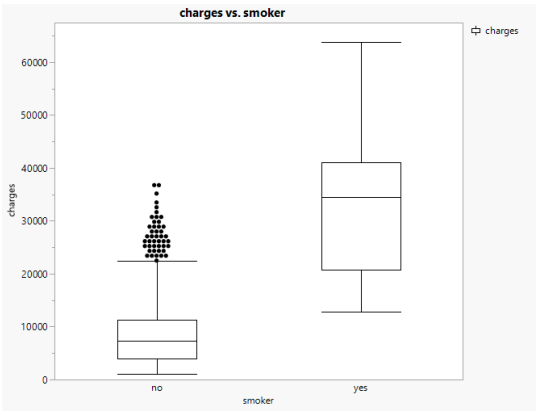
Smoker is the strongest categorical predictor. (See figure 2 in Appendix)

Graphs exploring potential relationships:

Histogram of charges

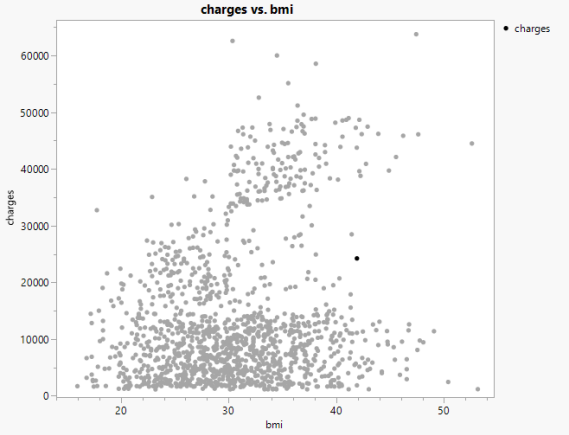


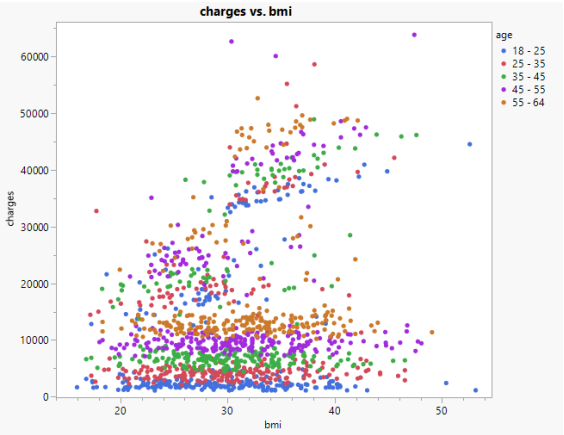
Boxplot smoker vs charges



Charges vs BMI scatterplot

charges vs smoker and age





Correlation & Multicollinearity Check

Correlation among quantitative predictors is weak (age–bmi ≈ 0.11 bmi–children ≈ 0.01 age–children ≈ 0.04).

Correlations

	age	bmi	children	charges
age	1.0000	0.1113	0.0439	0.2964
bmi	0.1113	1.0000	0.0125	0.2002
children	0.0439	0.0125	1.0000	0.0693
charges	0.2964	0.2002	0.0693	1.0000

The correlations are estimated by Row-wise method.

All VIF values were approximately 1, indicating no detectable multicollinearity. This means the predictors are not linearly redundant, and coefficient estimates are stable. This also validates the predictor set selected by Stepwise and All-Possible Regressions.

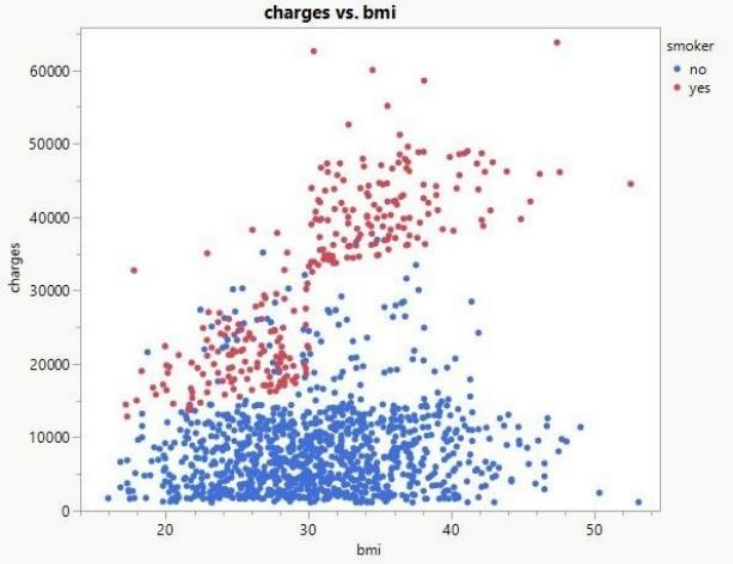
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	-213.5064	942.4549	-0.23	0.8208	-2062.366	1635.3537	.
age	256.85503	11.89585	21.59	<.0001*	233.51836	280.1917	1.0152045
bmi	323.74485	27.32314	11.85	<.0001*	270.14371	377.346	1.0126466
children	477.46406	137.5682	3.47	0.0005*	207.58971	747.33841	1.0020787
smoker[no]	-11925.8	205.4252	-58.05	<.0001*	-12328.8	-11522.81	1.0008868

Scatterplots & Interaction Evidence

Scatterplots reveal:

Smokers have much steeper BMI–charge slopes than non-smokers



Two clearly separated clusters appear when color-coding by smoker status

→ Strong visual evidence of a bmi× smoker interaction.

Methods

Study Design

This project uses an observational, cross-sectional dataset. Because the data were not collected experimentally, the analysis can identify associations among predictors and insurance charges but cannot establish causal relationships.

Variable Screening

Stepwise and all-possible-models screening identify:

- Retained: age, BMI, children, smoker
- Rejected: sex, region (no improvement in AICc, BIC, or R^2)

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	-213.50638	1	0	0.000	1
☐	☑	age	256.855032	1	1.71e+10	466.214	7e-89
☐	☐	sex{female-male}	0	1	4294573	0.117	0.73217
☐	☑	bmi	323.744854	1	5.142e+9	140.393	7.3e-31
☐	☑	children	477.46406	1	4.412e+8	12.046	0.00054
☐	☑	smoker{no-yes}	-11925.805	1	1.23e+11	3370.294	0
☐	☐	region{northwest&southwest&northeast-southeast}	0	1	82679699	2.260	0.13303
☐	☐	region{northwest&southwest-northeast}	0	1	94989373	2.597	0.10733
☐	☐	region{northwest-southwest}	0	1	52589692	1.436	0.23095

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	smoker{no-yes}	Entered	0.0000	1.21e+11	0.6210	695.36	2	27600	27615.5 ◯
2	age	Entered	0.0000	1.98e+10	0.7221	156.77	3	27187.7	27208.5 ◯
3	bmi	Entered	0.0000	5.165e+9	0.7485	17.491	4	27056.4	27082.3 ◯
4	children	Entered	0.0005	4.412e+8	0.7508	7.4233	5	27046.3	27077.5 ◉

All Possible Models

Ordered up to best 4 models up to 8 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC
smoker{no-yes}	1	0.6210	7454.90	27600.0	27615.5 ●
age	1	0.0878	11564.7	28772.3	28787.9 ○
bmi	1	0.0401	11863.6	28840.5	28856.0 ○
region{northwest&southwest&northeast-southeast}	1	0.0056	12074.7	28887.6	28903.1 ○
age,smoker{no-yes}	2	0.7221	6385.92	27187.7	27208.5 ●
bmi,smoker{no-yes}	2	0.6598	7065.66	27457.8	27478.6 ○
children,smoker{no-yes}	2	0.6249	7418.89	27588.1	27608.8 ○
smoker{no-yes},region{northwest&southwest&northeast-southeast}	2	0.6214	7453.56	27600.5	27621.3 ○
age,bmi,smoker{no-yes}	3	0.7485	6077.00	27056.4	27082.3 ●
age,children,smoker{no-yes}	3	0.7245	6360.95	27178.3	27204.2 ○
age,smoker{no-yes},region{northwest&southwest&northeast-southeast}	3	0.7226	6382.04	27187.1	27213.1 ○
age,smoker{no-yes},region{northwest&southwest-northeast}	3	0.7223	6385.25	27188.5	27214.4 ○
age,bmi,children,smoker{no-yes}	4	0.7508	6051.94	27046.3	27077.5 ●
age,bmi,smoker{no-yes},region{northwest&southwest&northeast-southeast}	4	0.7490	6073.49	27055.8	27086.9 ○
age,bmi,smoker{no-yes},region{northwest-southwest}	4	0.7488	6076.17	27057.0	27088.1 ○
age,bmi,smoker{no-yes},region{northwest&southwest-northeast}	4	0.7488	6076.18	27057.0	27088.1 ○
age,bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast}	5	0.7512	6049.07	27046.1	27082.4 ●
age,bmi,children,smoker{no-yes},region{northwest&southwest-northeast}	5	0.7511	6050.33	27046.6	27082.9 ○
age,bmi,children,smoker{no-yes},region{northwest-southwest}	5	0.7510	6051.15	27047.0	27083.3 ○
age,sex{female-male},bmi,children,smoker{no-yes}	5	0.7508	6053.94	27048.2	27084.5 ○
age,bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest&southwest-northeast}	6	0.7517	6045.44	27045.5	27087.0 ●
age,bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest-southwest}	6	0.7515	6048.10	27046.7	27088.1 ○
age,bmi,children,smoker{no-yes},region{northwest&southwest-northeast},region{northwest-southwest}	6	0.7513	6049.56	27047.3	27088.8 ○
age,sex{female-male},bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast}	6	0.7512	6051.08	27048.0	27089.5 ○
age,bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest&southwest-northeast},region{northwest-southwest}	7	0.7519	6044.44	27046.1	27092.7 ●
age,sex{female-male},bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest&southwest-northeast}	7	0.7517	6047.43	27047.4	27094.0 ○
age,sex{female-male},bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest-southwest}	7	0.7515	6050.10	27048.6	27095.2 ○
age,sex{female-male},bmi,children,smoker{no-yes},region{northwest&southwest-northeast},region{northwest-southwest}	7	0.7514	6051.56	27049.2	27095.9 ○
age,sex{female-male},bmi,children,smoker{no-yes},region{northwest&southwest&northeast-southeast},region{northwest&southwest-northeast},region{northwest-southwest}	8	0.7520	6046.43	27048.0	27099.8 ●

The 4-predictor model : age + bmi + children + smoker(no-yes) outperforms the 5-predictor model in parsimony without meaningful loss in accuracy.

Model Building Strategy

A sequence of nested multiple regression models was fit to evaluate the research questions and improve model adequacy:

Model 1 (Main Effects)

$$E(y)= \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4$$

Where x_1= age, x_2 = bmi, x_3 = children, x_4 = smoker[no]

Model 2 (Add BMI * smoker Interaction Model)

$$E(y)=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4 + \beta_5x_2x_4$$

Where x_2x_4 is the interaction term

Added the BMI × smoker interaction to test whether the effect of BMI differs between smokers and non-smokers — the primary research question.

Model 3 (Log-Transformed Model)

$$\ln(y)= \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4 + \beta_5x_2x_4$$

Modeled ln(charges) to correct the severe heteroscedasticity and non-normality present in Models 1 and 2.

The Box–Cox procedure indicated $\lambda \approx 0$, justifying the natural log transformation.

Model Comparison and Selection

Diagnostics performed:

- Global F-tests and t-tests

- Residual vs. predicted plots (linearity & homoscedasticity)
- Q–Q plots and histograms (normality)
- Shapiro–Wilk tests
- Outlier and leverage diagnostics (studentized residuals, hat values)
- Influence statistics (Cook's D, external studentized residuals)

Validation methods:

- Stepwise **training vs. validation R^2 curves**
- **80/20 train–test split** (RSquare and RASE comparison)
- **PRESS statistic** for final model validation

All statistical analyses were performed in JMP. Detailed graphs, tables, and numerical output (e.g., distribution plots, interaction plots, residual diagnostics, influence measures, and validation charts) are provided in the Appendix titled Final Project EDA and Model Development: Feifei Li.

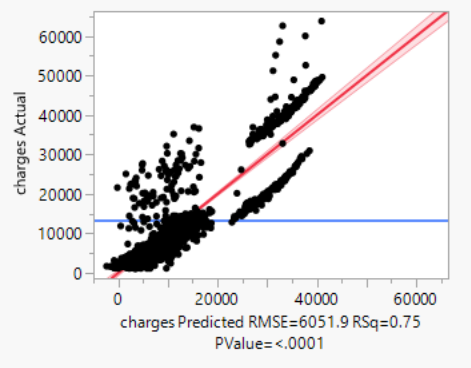
Results

Model 1 – Main Effects

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	11712.298	988.8939	11.84	<.0001*	9772.3365	13652.26
age	256.85503	11.89585	21.59	<.0001*	233.51836	280.1917
bmi	323.74485	27.32314	11.85	<.0001*	270.14371	377.346
children	477.46406	137.5682	3.47	0.0005*	207.58971	747.33841
smoker[no]	-23851.61	410.8505	-58.05	<.0001*	-24657.6	-23045.62

$$E(y) = 11712.298 + 256.85503x_1 + 323.74485x_2 + 477.46406x_3 - 23851.61x_4$$



Key Findings

All predictors are statistically significant ($p < .001$).

Smokers: + **$23,852** average higher charges than non-smokers.

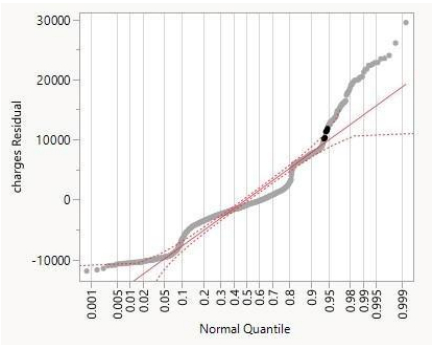
Adjusted R^2 = **0.7508** RMSE ≈ **6052** → prediction error ≈ ±12,100

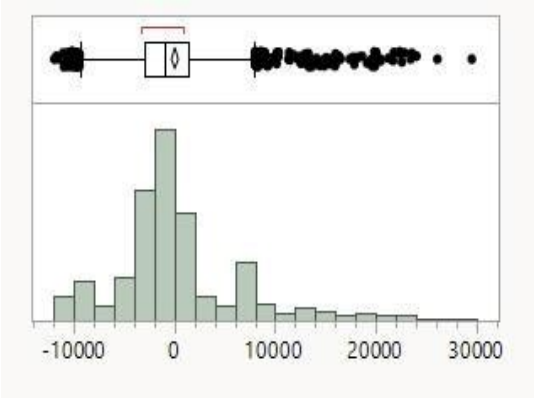
CV=6051.94/13255.38×100≈**45.66%(poor prediction accuracy)**

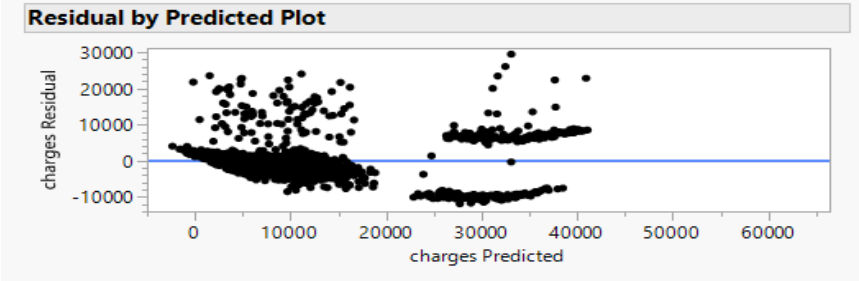
Summary of Fit	
RSquare	0.750763
RSquare Adj	0.750013
Root Mean Square Error	6051.935
Mean of Response	13255.38
Observations (or Sum Wgts)	1335

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1.4673e+11	3.668e+10	1001.569
Error	1330	4.8712e+10	36625921	Prob > F
C. Total	1334	1.9545e+11		<.0001*

Residuals show **severe heteroscedasticity, skewness, and curvature**





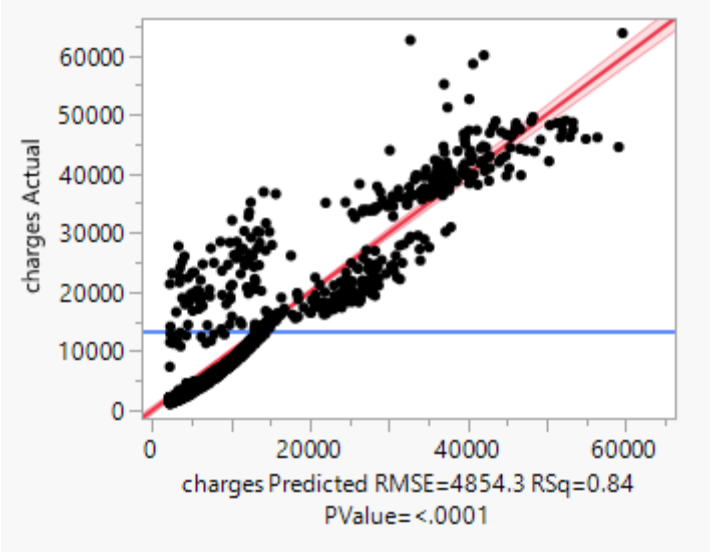


Residual diagnostics revealed a clear separation between smokers and non-smokers in the residual–predicted plot, forming **two distinct bands.** This indicates that the main-effects model does not adequately capture the relationship between BMI and charges, and suggests that the effect of BMI differs for smokers and non-smokers. Therefore**, an interaction term** is likely required to correctly specify the model.

Conclusion: Model 1 is misspecified and cannot capture differences in BMI effect between smokers and non-smokers.

Model 2 – Adding BMI × Smoker Interaction

$$E(y) = -22847.84+264.01x_1+1438.69x_2+513.61x_3+20065.29x_4 - 1430.67\mathbf{x_2x_4}$$



Key findings:

1. Interaction term BMI main effect = -1430.67 ($p < .0001$) → strongly significant

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-22847.84	1499.047	-15.24	<.0001*	-25788.6	-19907.09
age	264.00987	9.545404	27.66	<.0001*	245.28417	282.73557
bmi	1438.6853	46.52146	30.93	<.0001*	1347.4218	1529.9488
children	513.60664	110.3528	4.65	<.0001*	297.12202	730.09126
smoker[no]	20065.293	1649.626	12.16	<.0001*	16829.137	23301.448
bmi*smoker[no]	-1430.67	52.65624	-27.17	<.0001*	-1533.969	-1327.372

A **Partial F-Test** comparing Model 1 produced F = 738.21 with $p \approx 1.18 \times 10^{-129}$, providing overwhelmingly strong evidence that the **interaction improves prediction**. Because only one parameter (β_5) was added, this result agrees with the t-test and confirms that the interaction must be retained.

Custom Test

Parameter	
Intercept	0
age	0
bmi	0
children	0
smoker[no]	0
bmi*smoker[no]	1
=	0
Value	-715.3352394
Std Error	26.328117836
t Ratio	-27.17001055
Prob>|t|	1.1816e-129
SS	17395436113
Sum of Squares	17395436113
Numerator DF	1
F Ratio	738.20947332
Prob > F	1.1816e-129

2. **Interaction: Smokers vs. Non-Smokers**

   smoker = yes is the baseline ($x_4 = 0$), non-smoker = no is the indicator ($x_4 = 1$).

   - **For smokers (baseline): BMI slope = 1438.69**
   - **For non-smokers: BMI slope = 1438.69 − 1430.67 ≈ 8.02**

   BMI strongly influences smokers' charges but barely affects non-smokers.

3. Adjusted R^2 = **0.8392** (large improvement) RMSE ≈ **4854** CV ≈ **36.6%** (still far too high)

Summary of Fit

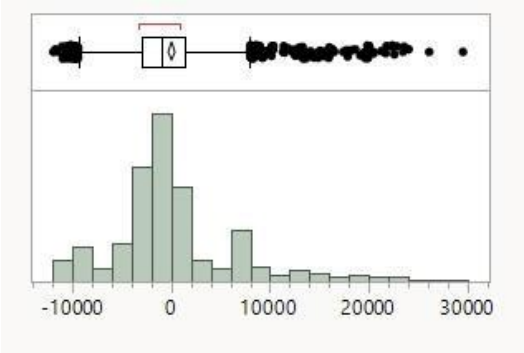
RSquare	0.839766
RSquare Adj	0.839164
Root Mean Square Error	4854.314
Mean of Response	13255.38
Observations (or Sum Wgts)	1335

Analysis of Variance

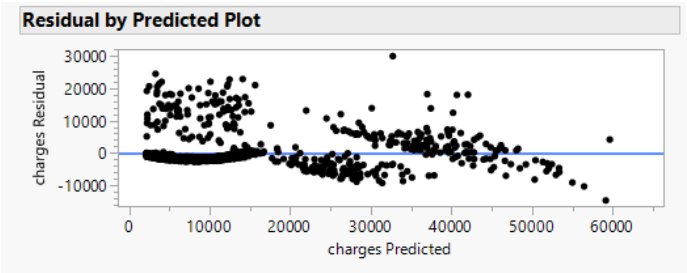
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	1.6413e+11	3.283e+10	1393.028
Error	1329	3.1317e+10	23564363	Prob > F
C. Total	1334	1.9545e+11		<.0001*

4. Remaining Assumption Violations

Residuals remain right-skewed



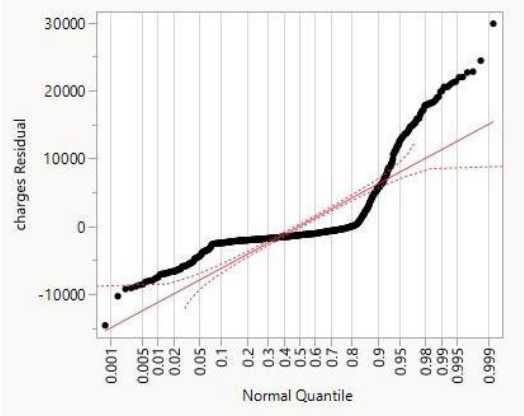
Heteroscedasticity is reduced but still evident



±3RMSE = 14.562.9 ,

It shows **potential outliers**

Q-Q plot shows **curveture**, it indicates violated the normality assumptions.



Shapiro–Wilk p < .0001 → reject normality , non-normal errors

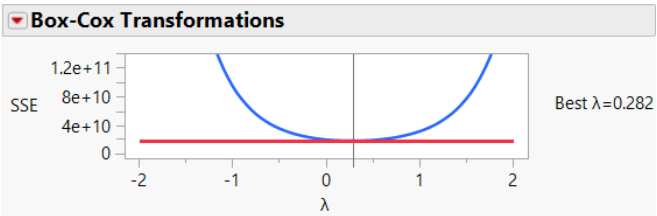
Goodness-of-Fit Test

	W	Prob<W
Shapiro-Wilk	0.9005086	<.0001*

	A^2	Simulated p-Value
Anderson-Darling	43.408222	<.0001*

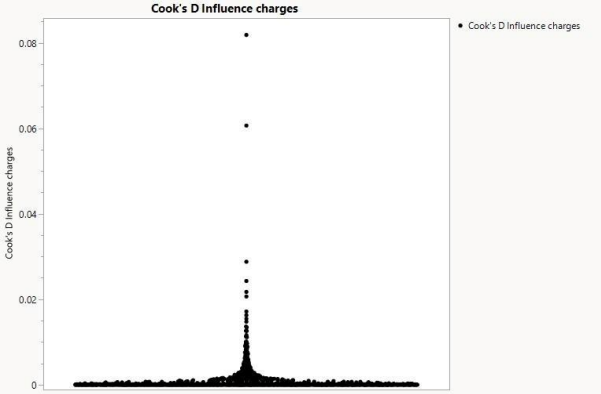
Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.

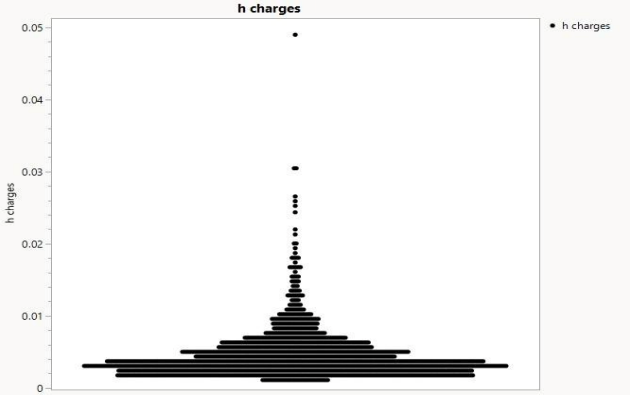
Box–Cox indicates λ ≈ 0 → natural log transform required



Cook's Distance and leverage plots show no influential observations

- D >F0.5 (df1 = 5, df2 =1330) = 0.871 or D >1 --> Influential
- Hat value h >2(6)/1335 = 0.009 , there is No high-leverage points





Conclusion: Model 2 diagnostics

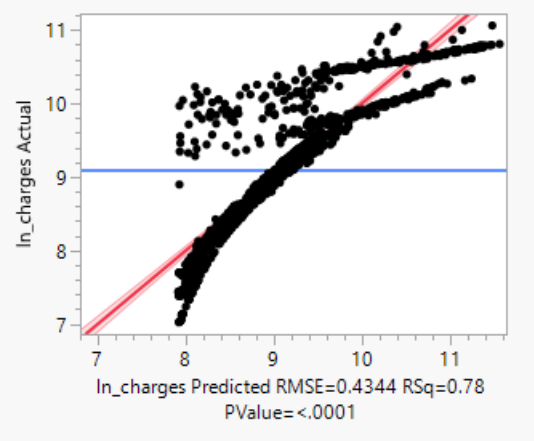
- Better than Model 1
- Still heteroscedastic
- Still non-normal (Shapiro–Wilk < .0001)
- Box–Cox indicates λ ≈ 0 → log transform required

Model 3 – Log-Transformed Model

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	7.465508	0.134132	55.66	<.0001*	7.2023738	7.7286422
age	0.0349422	0.000854	40.91	<.0001*	0.0332666	0.0366177
bmi	0.0449857	0.004163	10.81	<.0001*	0.0368196	0.0531519
children	0.1021047	0.009874	10.34	<.0001*	0.082734	0.1214754
smoker[no]	-0.197061	0.147606	-1.34	0.1821	-0.486627	0.0925048
bmi*smoker[no]	-0.043942	0.004712	-9.33	<.0001*	-0.053185	-0.034699

$$\ln(y) = 7.4655 + 0.03494x_1 + 0.04499x_2 + 0.1021x_3 - 0.1971x_4 - 0.0439x_2x_4$$



x_4 = smoker[no](so **smoker = yes → $x_4 = 0$ (baseline), non-smoker = no → $x_4 = 1$**)

These coefficients reflect percentage changes:

β5=-0.0439→ about **−4.3 percentage points change in the BMI effect when you move from smoker → for non-smokers, the BMI slope is 4.3 percentage points smaller than for smokers.**

Predictor	Interpretation
age	+3.5% charges per year
bmi	+4.5% per bmi unit
children	+10.2% per child
smoker (main effect)	Not significant after interaction
bmi × smoker	Very strong effect

Smokers: BMI increases charges ~4.5% per unit

Non-smokers: BMI increases charges only ~0.1% per unit

Key findings

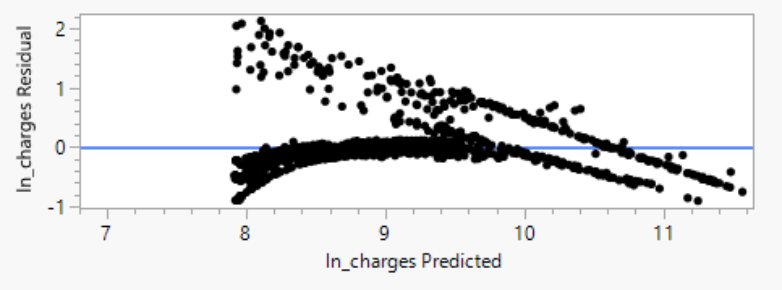
1. Adjusted R^2 = **0.776** (log scale) RMSE = **0.434** CV = **4.77% → Excellent**

Summary of Fit	
RSquare	0.777158
RSquare Adj	0.77632
Root Mean Square Error	0.434356
Mean of Response	9.098084
Observations (or Sum Wgts)	1335

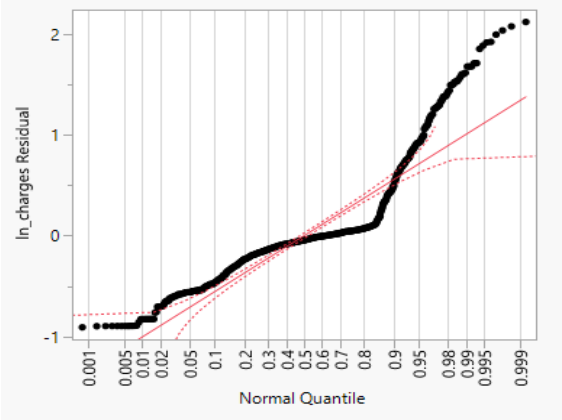
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	874.4412	174.888	926.9757
Error	1329	250.7363	0.189	Prob > F
C. Total	1334	1125.1776		<.0001*

The CV dropped from 45.6% (Model 1) and 36.6% (Model 2) down to 4.77%, showing that the log transform dramatically improved predictive quality.

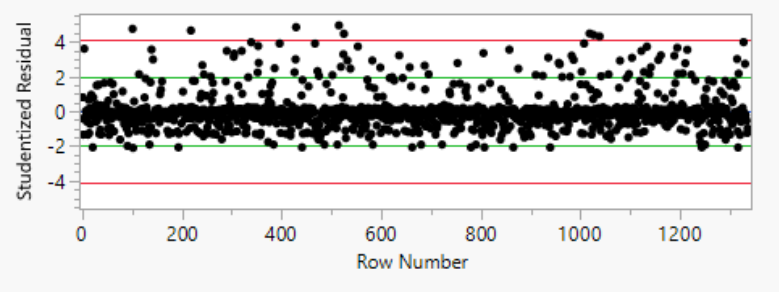
2. Residuals: substantially improved



3. Q–Q plot: mild tail deviations only



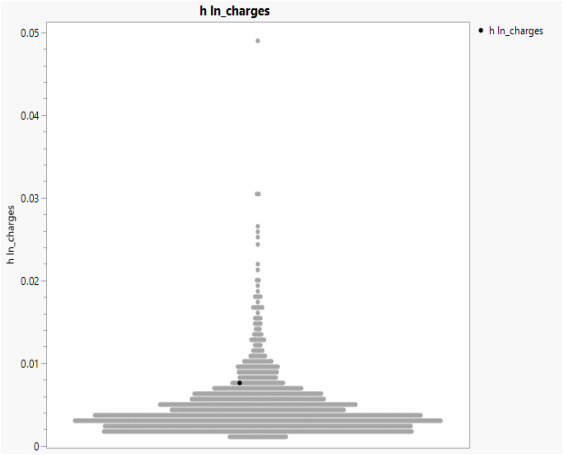
4. Small groups of natural outliers remain, expected in healthcare cost data



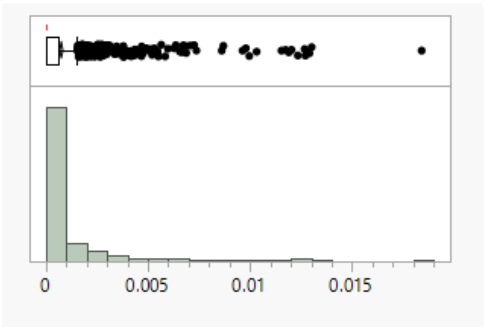
5. No high value leverage points.

   Average hat vaule (h = (k+1)/n =0.00449) k=5 n = 1335

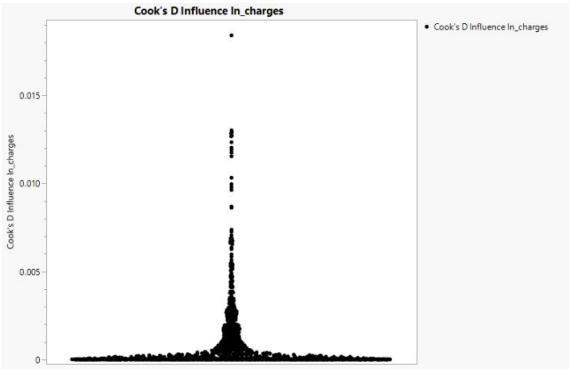
   Rule of Thumb: Hat value h >2(6)/1335 = 0.009



6. No influential observations (Cook's D < threshold)



Quantiles		
100.0%	maximum	0.01841018
99.5%		0.012448
97.5%		0.0057639
90.0%		0.00236951
75.0%	quartile	0.00061532
50.0%	median	3.79227e-5
25.0%	quartile	4.84216e-6
10.0%		5.39793e-7
2.5%		2.85141e-8
0.5%		5.3412e-10
0.0%	minimum	1.7369e-12



If D >F0.5 (df1 = 5, df2 =1330)= 0.871

or D >1 → Influential

Applying a natural log transformation to charges substantially reduced the major assumption violations found in Models 1 and 2. Residual variance stabilized, the distribution became much more symmetric, and Q–Q plots showed only mild tail deviations—patterns that are common and expected in healthcare cost data. No high-leverage or influential observations were detected. Predictive accuracy improved dramatically (CV = 4.77%), and the BMI × smoker interaction remained statistically strong.

Model 3 is statistically valid, interpretable, and represents the best-performing model among those evaluated. These findings directly support the significance of the research question by showing how lifestyle factors—particularly the combination of smoking and high BMI—drive large differences in medical costs that are relevant to insurers, policymakers, and consumers.

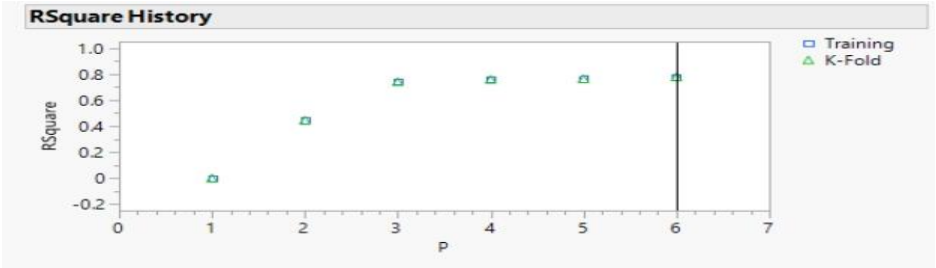
Although assumption adherence improved substantially, Model 3 still shows mild curvature in the Actual vs Predicted plot and some uneven residual spread due to natural clustering (e.g., smokers vs non-smokers). These remaining patterns are typical for healthcare expenditure data and do not meaningfully diminish the model's strong predictive performance.

Model Validation

1. Stepwise validation: training and validation R^2 align closely

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	RSquare K-Fold
1	All	Entered	.	.	0.7772	6	6	1570.15	1606.44	0.7756
2	bmi*smoker{no-yes}	Removed	0.0000	16.41022	0.7626	90.981	5	1652.76	1683.88	0.7610
3	bmi	Removed	0.0000	5.660221	0.7575	118.98	4	1678.73	1704.67	0.7565
4	children	Removed	0.0000	19.90177	0.7399	222.47	3	1770.72	1791.48	0.7386
5	age	Removed	0.0000	333.2444	0.4437	1986.8	2	2783.44	2799.01	0.4428
6	smoker{no-yes}	Removed	0.0000	499.2246	-0.000	4630.9	1	3564.3	3574.69	-0.001
7	Best	Specific	.	.	0.7772	6	6	1570.15	1606.44	0.7756



2. 80/20 split: validation RSquare & RASE ≈ training → Good generalization

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.7900	0.42051	1076
Validation Set	0.7216	0.48452	259

3. PRESS R^2 = **0.7752**, nearly identical to ordinary R^2

Press

Residual	SSE	RMSE	RSquare
Press	252.98463588	0.43531781	0.7752
Ordinary	250.73633005	0.43435632	0.7772

Across all validation techniques—stepwise validation, train–test splitting, and PRESS—Model 3 shows strong predictive reliability and minimal overfitting. These results confirm that the log-transformed BMI × smoker interaction model is robust, stable, and appropriate as the final model.

Metric	Model 1: Main Effects	Model 2: Interaction	Model 3: Log-Transformed
Equation Form	Linear	Linear + BMI×Smoker	Log-linear + BMI×Smoker
Adjusted R^2	0.7508	0.8392	0.776 (log scale)
RMSE	6052	4854	0.434
CV (Prediction Error)	**45.66%**	**36.6%**	**4.77%**
Skewness & Normality	Severe skewness, non-normal	Still skewed, non-normal	Nearly normal residuals
Constant Variance	Violated	Violated	Greatly improved
Outliers/Influential	None severe	None	None
Interpretation Quality	Moderate	Strong	Strongest (percent changes)
Overall Fit	Poor	Better	Best
Final Choice?	No	No	**Yes**

Final Model Selection

After evaluating three nested regression models, the log-transformed model with the BMI × smoker interaction (Model 3) provides the best overall performance among our candidate models. Model 1 identified significant predictors but showed severe heteroscedasticity, strong right skewness, and a two-cluster residual pattern separating smokers and non-smokers, indicating structural misspecification. Model 2 successfully corrected this by adding the BMI × smoker interaction, and both the partial F-test and t-test strongly supported retaining this term. However, substantial heteroscedasticity and non-normality remained.

Model 3 addressed these issues by applying a natural log transformation to the response variable, as suggested by the Box–Cox procedure ($\lambda \approx 0$). This transformation greatly stabilized the

variance, improved residual symmetry, and produced much more reliable Q–Q plots. Although Model 3 is not perfect—the Actual vs Predicted plot still shows mild curvature and the constant variance assumption is not fully satisfied due to natural clustering of observations—these patterns are typical and expected for healthcare expenditure data, which commonly contain subgroups with inherently different cost structures.

Model 3 retains a strong BMI × smoker interaction effect, offers clear percent-change interpretations, and demonstrates excellent predictive accuracy (CV ≈ 4.77%). Validation via stepwise training–validation curves, an 80/20 train–test split, and the PRESS statistic all show strong generalizability with no overfitting.

Therefore, Model 3 is the most appropriate and defensible final model given our project objectives.

If further improvements were desired beyond the scope of this project, potential extensions could include exploring nonlinear terms (e.g., Age^2 or BMI^2) or additional interactions to help account for the curvature in

Conclusion and Discussion

This study examined how demographic and lifestyle factors influence U.S. medical insurance charges, with a focus on whether BMI affects smokers and non-smokers differently. Smoking emerged as the strongest driver of charges, and the BMI × smoker interaction revealed a striking finding: BMI greatly increases costs for smokers but has only a modest effect for non-smokers.

Transforming charges using the natural log significantly improved model adequacy by stabilizing variance and reducing skewness. Model 3 offered the best balance of interpretability, statistical validity, and predictive accuracy. Although mild curvature and small subgroup-based variance differences remain, these patterns are typical in medical cost data and do not impair overall model performance.

Limitations

- Cross-sectional data cannot determine causality
- Charges are simulated or anonymized; real-world underwriting may include more predictors
- Nonlinear terms (Age^2, BMI^2) may capture remaining curvature

Implications

- Supports differential pricing for high-BMI smokers
- Highlights value of preventive smoking-cessation programs
- Helps explain cost drivers for insurers and policymakers

Model 3 successfully answers the research questions, demonstrating that lifestyle factors—especially smoking combined with high BMI—drive substantial differences in healthcare costs.

References

Metropolitan State University. (2025). *STAT 311: Regression Analysis (Fall 2025).*

Kaggle. *Health Insurance Charges Dataset.* Nalisha

https://www.kaggle.com/datasets/nalisha/health-insurance-charges-dataset.

Appendix

Final Project EDA and Model Development: Feifei Li (2025)

https://docs.google.com/document/d/1p5d_nwapZUGMHNCHDuholCZnV0oVv9Cox8AQX0oZLo4/edit?tab=t.0#heading=h.fr55k0j9zlnz

Github links for datasets and journal files created by JMP

https://github.com/fayfaycn/Stat-311-Final-Project-

Figure 1

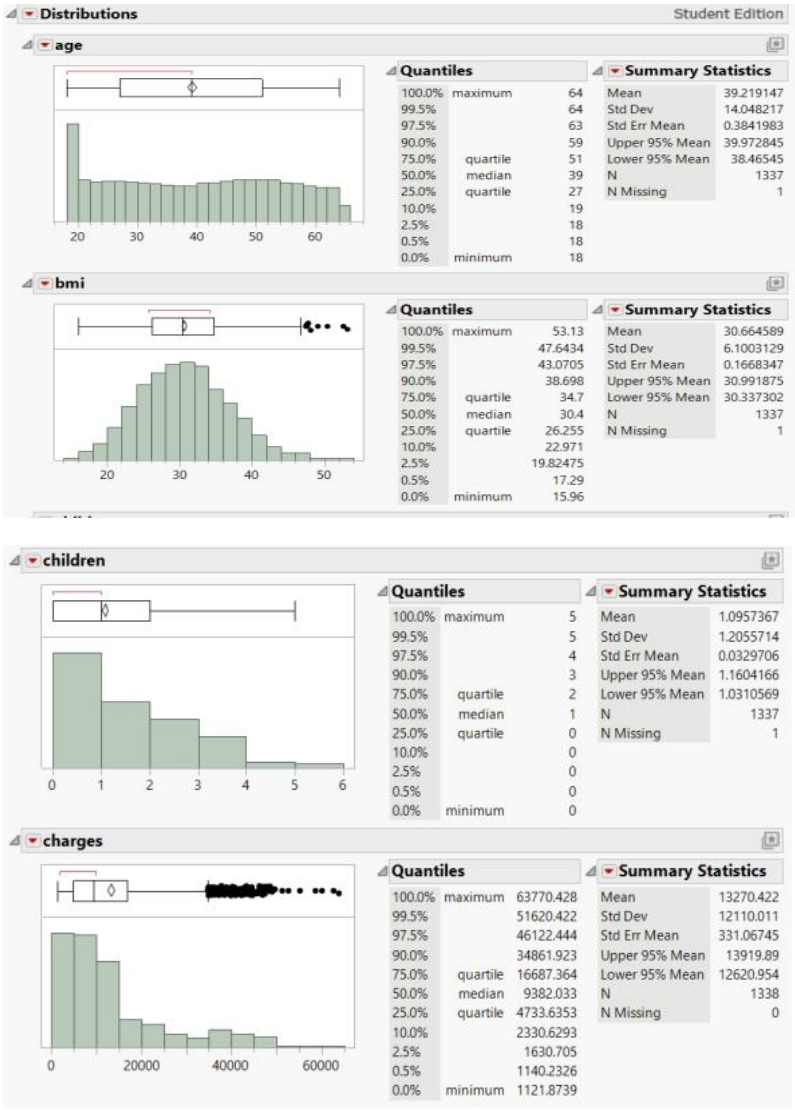
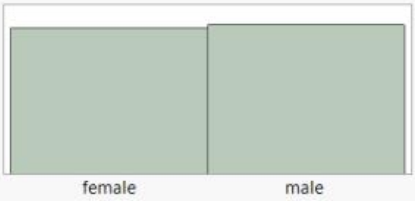

Distributions
Student Edition
age
Quantiles
100.0% maximum 64
99.5% 64
97.5% 63
90.0% 59
75.0% quartile 51
50.0% median 39
25.0% quartile 27
10.0% 19
2.5% 18
0.5% 18
0.0% minimum 18
Summary Statistics
Mean 39.219147
Std Dev 14.048217
Std Err Mean 0.3841983
Upper 95% Mean 39.972845
Lower 95% Mean 38.46545
N 1337
N Missing 1
20 30 40 50 60
bmi
Quantiles
100.0% maximum 53.13
99.5% 47.6434
97.5% 43.0705
90.0% 38.698
75.0% quartile 34.7
50.0% median 30.4
25.0% quartile 26.255
10.0% 22.971
2.5% 19.82475
0.5% 17.29
0.0% minimum 15.96
Summary Statistics
Mean 30.664589
Std Dev 6.1003129
Std Err Mean 0.1668347
Upper 95% Mean 30.991875
Lower 95% Mean 30.337302
N 1337
N Missing 1
20 30 40 50
children
Quantiles
100.0% maximum 5
99.5% 5
97.5% 4
90.0% 3
75.0% quartile 2
50.0% median 1
25.0% quartile 0
10.0% 0
2.5% 0
0.5% 0
0.0% minimum 0
Summary Statistics
Mean 1.0957367
Std Dev 1.2055714
Std Err Mean 0.0329706
Upper 95% Mean 1.1604166
Lower 95% Mean 1.0310569
N 1337
N Missing 1
0 1 2 3 4 5 6
charges
Quantiles
100.0% maximum 63770.428
99.5% 51620.422
97.5% 46122.444
90.0% 34861.923
75.0% quartile 16687.364
50.0% median 9382.033
25.0% quartile 4733.6353
10.0% 2330.6293
2.5% 1630.705
0.5% 1140.2326
0.0% minimum 1121.8739
Summary Statistics
Mean 13270.422
Std Dev 12110.011
Std Err Mean 331.06745
Upper 95% Mean 13919.89
Lower 95% Mean 12620.954
N 1338
N Missing 0
0 20000 40000 60000

Figure 2

Distributions

Student Edition

sex



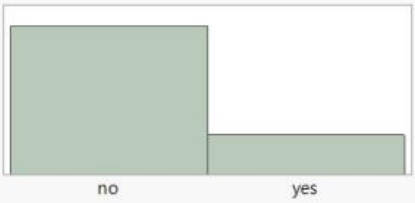
Frequencies

Level	Count	Prob
female	662	0.49477
male	676	0.50523
Total	1338	1.00000

N Missing 0

2 Levels

smoker



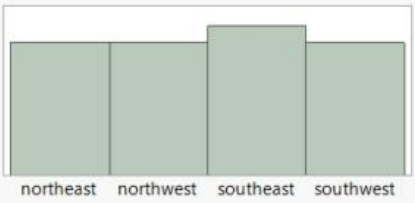
Frequencies

Level	Count	Prob
no	1064	0.79522
yes	274	0.20478
Total	1338	1.00000

N Missing 0

2 Levels

region



Frequencies

Level	Count	Prob
northeast	324	0.24215
northwest	325	0.24290
southeast	364	0.27205
southwest	325	0.24290
Total	1338	1.00000

N Missing 0

4 Levels