An Analysis of Insurance Data Set Group 17

Poojit Tummalapalli Vandna Hrushikesh Thanikonda

STEVENS INSTITUTE OF TECHNOLOGY Statistical Methods

Contents

1. Introduction

1.1 Data Description

Data Source: https://www.kaggle.com/datasets/nazeernazeer/insurance-dataset-for-statistical-analysis

The insurance dataset used by our team contains 1,338 policyholder records with seven core variables per row: age, sex, body mass index (BMI), number of children, smoker status, geographic region, and individual insurance charges. The data had no missing entries or duplicate rows originally.

The age range for policyholders is 18 to 64 years, with a mean of 39.2 and a standard deviation of 14.1. The distribution was slightly right-skewed, meaning there was a heavier concentration of younger adults and a tapering tail toward the upper bound of 64. This skew suggested older policyholders were relatively less common in this dataset which could influence risk modeling if age interacted nonlinearly with other factors.

The gender column is nearly balanced with a 51% male and 49% female ratio. Such parity ensured that modeling efforts could fairly assess sex-based risk differences without major imbalances. It also increased the generalizability of findings across male and female segments since both groups are nearly equal.

BMI values ranged from 15.9 to 53.1, with an average of 30.7 and a standard deviation of 6.1. This range of values captured both underweight individuals and reasonably overweight individuals. The distribution was right skewed because of a subset of high-BMI policyholders which highlighted a potential nonlinear relationship between BMI and healthcare costs. This seemed reasonable given the researched association between obesity and higher medical risk.

Household composition, as measured by the number of children covered, had a range from 0 to 5 with a mode of 0 as 36% of data subjects had no dependents. Since approximately one-third of the dataset consisted of single-adult policies, family structure could cause an increase of risk and cost patterns, especially when combined with other demographic factors.

Smoker status showed that 20% of policyholders were smokers. Since smoking is a known major health risk, this subgroup was expected to drive medical costs higher. In fact, the inclusion of smoking status was critical for any predictive model of costs, as smokers often incurred frequent and expensive medical claims.

The dataset was split evenly across four regions, namely northeast, northwest, southeast, and southwest, which minimized regional bias. This uniformity allowed for more reliable comparisons of regional cost differentials and ensured that no single area had disproportionately influenced our testing.

Finally, insurance charges themselves ranged from $1,121.87 to $63,770.43, with a mean of $13,270.42 and a standard deviation of $12,105.49. The charge distribution was markedly right-skewed meaning a small subset of high-cost claims had increased the upper tail. This

fact underscored the importance of considering robust modeling techniques or transformations when predicting or analyzing claim amounts.

1.2 Software and Variables

We analyzed the dataset using Python and its built-in libraries: Pandas, Numpy, Seaborn, SciPy, Statsmodel and Sklearn for implementing different machine learning models.

List of Variables:

μ_x - Population Mean

$\mu_{\bar{x}}$ - Sample Mean

$\sigma_{\bar{x}}$ - Standard Deviation of Sample Means

$\sigma_x/\sqrt{n}$ – Standard Deviation of population of sample means

H_o - Null Hypothesis

H_a - Alternate Hypothesis

1.3 Goal

Our goal is first to analyze the insurance dataset to find relationships among different variables and variations of the target variable and second implement different predictive models like linear regression, SVM, K-means and random forest regressor to predict medical charges.

2. Analysis and Visualization of Data

Histograms made on the age variable show a right-skewed distribution as the majority of policyholders are clustered between the ages of 25 and 55. The kernel density estimate overlaid on the histogram smoothed out minor fluctuations, confirming that the distribution peaked just below the mean (39.2 years) and then tapered gradually. The skewness of the age distribution hovered around 0.4 which indicates mild asymmetry while the kurtosis was close to 3.1 meaning the distribution was marginally more peaked than the Gaussian distribution.

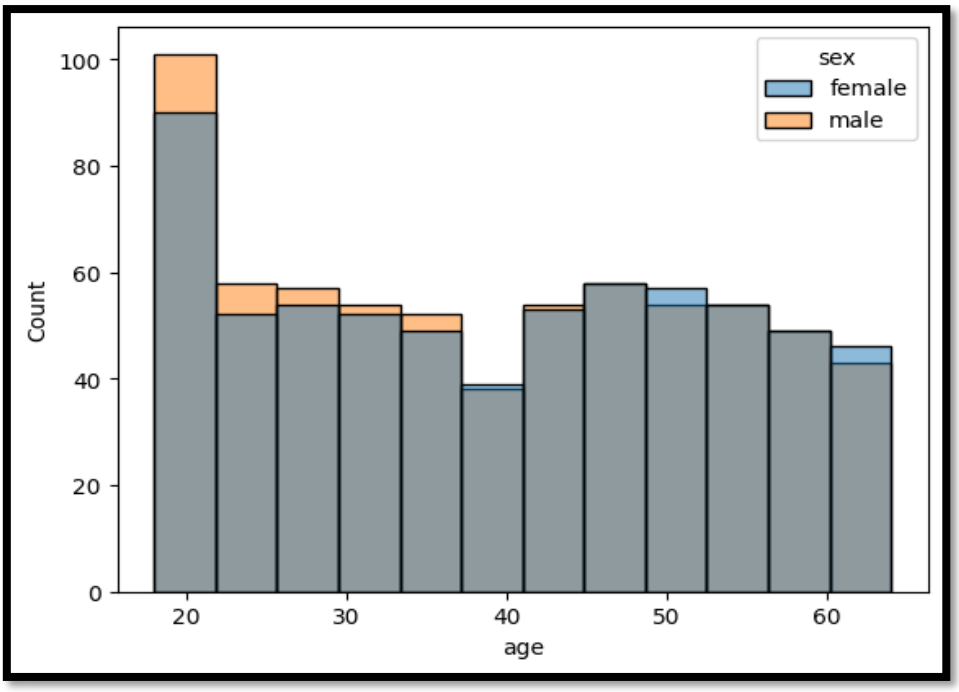


Figure 1 Age Distribution

In contrast, the BMI distribution appears remarkably symmetric but leptokurtic. The histogram bars rose sharply around the 30–32 BMI range and the density curve formed a tall, narrow peak there. The shoulders of the distribution fell off quickly toward both underweight and obese extremes, producing heavy tails relative to a normal curve. Numerically, BMI skewness was essentially zero (–0.02), but kurtosis measured about 4.5, signaling that more observations lay in the center and in the tails than one would expect under strict normality.

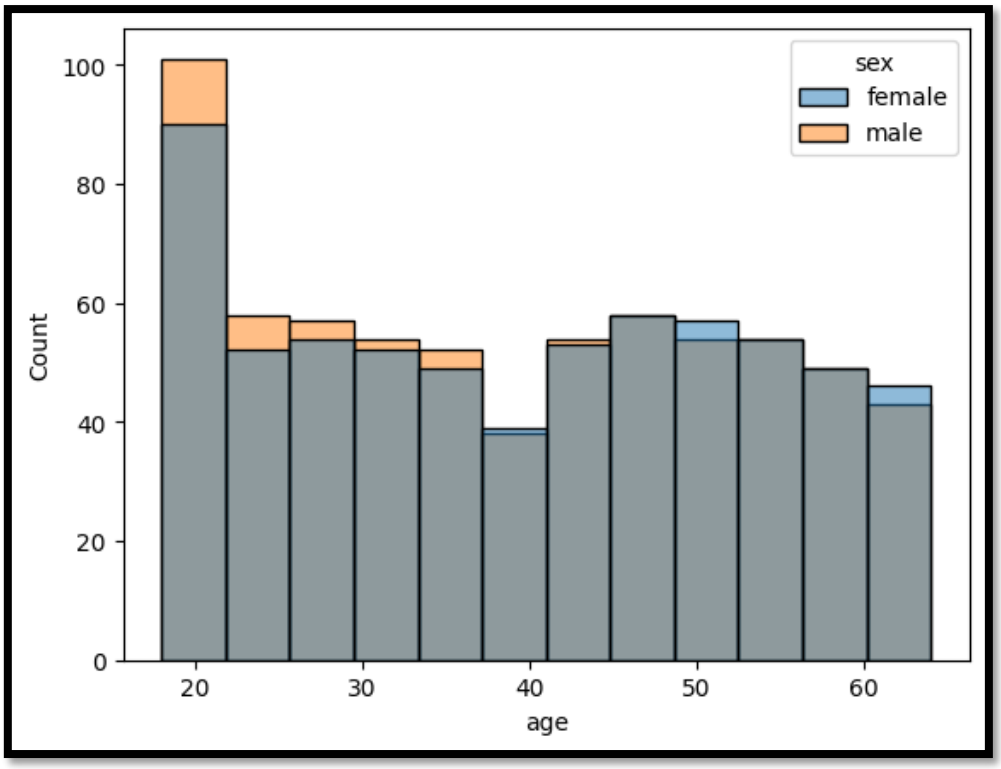


Figure 2 BMI Distribution

The raw insurance charges histogram had a higher rate of change as a heavy right skew was apparent, with most policyholders incurring charges between $3,000 and $15,000, but there still existed a tail of high-cost claims stretching past $50,000. The density plot confirmed this asymmetry, rising sharply at lower charges before decaying slowly through the upper extremes. Raw charge skewness clocked in at roughly 3.5 and kurtosis at about 18, highlighting the extreme outliers in medical costs.

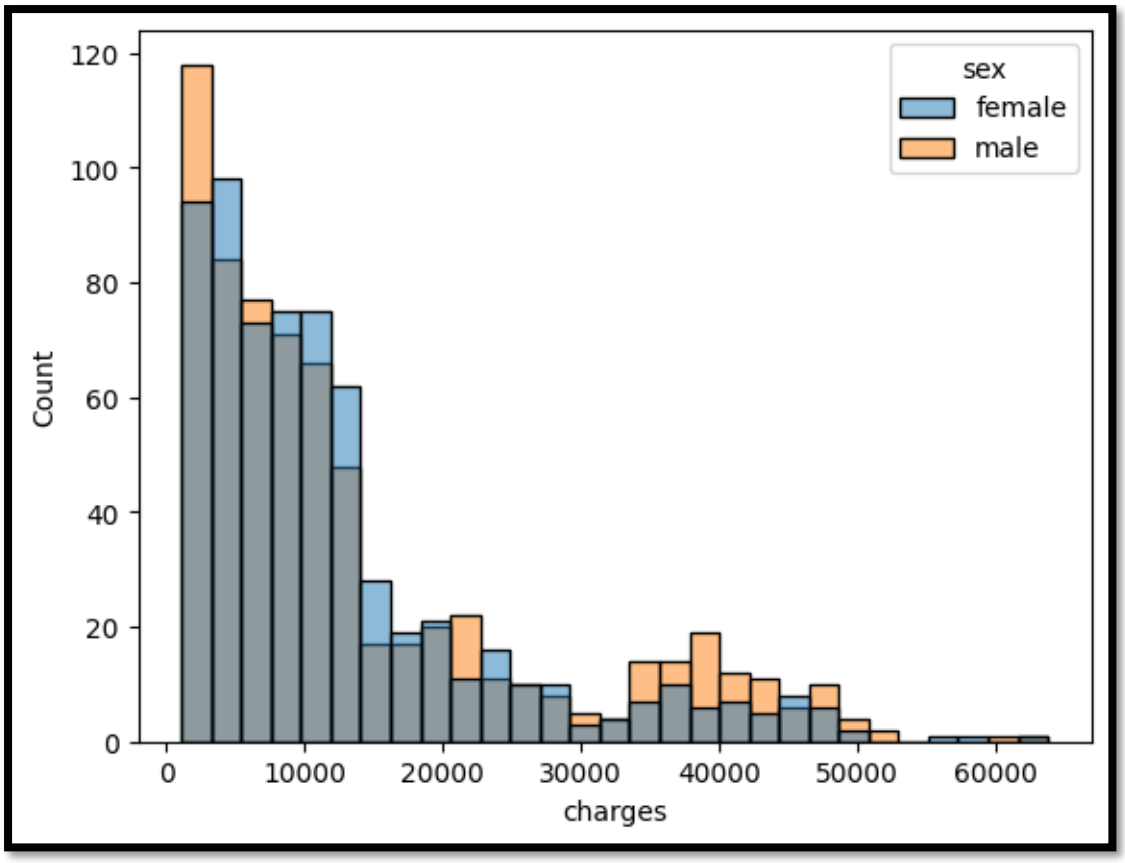


Figure 3 Charges Distribution

After applying a natural log-transformation to charges, the histogram contracted into a much more symmetric, bell-shaped form. The corresponding density curve matched a near-normal pattern, centered around log-charge ≈ 9.5 (which corresponds to a medical charge of about $13,500). Skewness dropped to 0.6 and kurtosis dropped to 4, which is closer to the Gaussian ideal making log-charges far more suitable for linear modeling assumptions. QQ-plots showed log-charges and BMI fell almost entirely along the 45° reference line, with slight deviations at the extremes, whereas charges diverged sharply in the upper tail. The age QQ-plot had shown modest deviation above the 75th percentile, consistent with its mild skew.

Finally, the discrete bar plot of the number-of-children variable had revealed a clear modal frequency at zero: about 36% of policyholders had no dependents. Frequencies then tapered steadily through one, two, three, and up to five children, with fewer than 2% of records at the upper end. The pattern suggested a right-skewed count distribution, reinforcing that most households were either single adults or small families.

We noticed that the average number of children in a family is 1 child according to the boxplot.

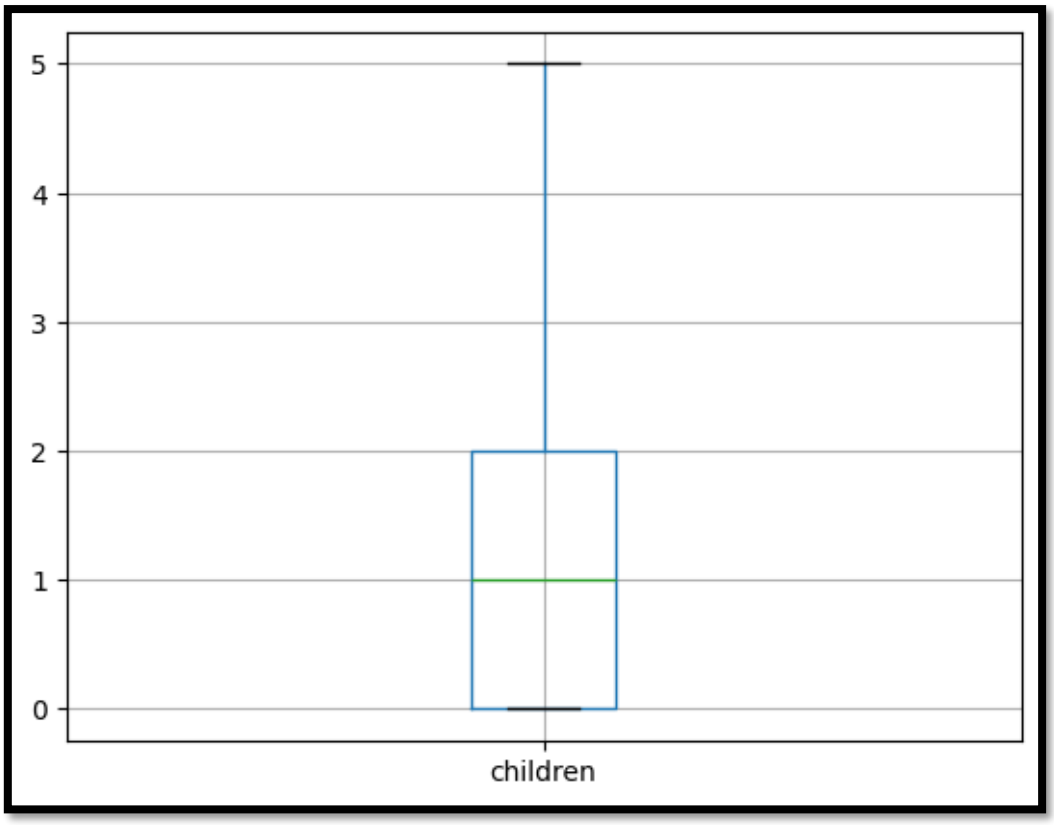


Figure 4 Boxplot of Children

The density estimation of the age distribution appeared uniform across the 18–64 range meaning no particular age band had dominated the sample. The age curve ran almost flat from the early twenties through the fifties and dipped only slightly among the oldest policyholders. By contrast, the BMI density plot peaked sharply around 29–31 and demonstrated that a large proportion of policyholders are classified as overweight ($BMI \geq 25$). A long right-hand tail extended beyond a BMI of 40, illustrating that a minority of policyholders had extreme obesity.

Boxplots on the smoking status column underscored the financial burden of tobacco use. Smokers had a median charge of about $23,000, which was nearly four times the $6,000 median for non-smokers. Their interquartile range spanned roughly $10,000 to $40,000 while non-smokers' ranged from $3,500 to $9,500. Moreover, the upper whisker for smokers extended beyond $60,000, and several extreme outliers climbed above $63,000—none of which appeared in the non-smoker group. These visuals confirmed that smoking not only raised typical costs but also introduced the greatest unpredictability in high-cost claims.

Regional box plots showed modest geographic variation. The Northeast had the highest median charge at approximately $14,500, compared with about $14,000 in the Southwest. However, the interquartile ranges overlapped substantially across all four regions, indicating that regional factors shifted central tendencies only slightly. The Northeast and Southeast displayed the most pronounced upper-tail extensions, which suggested that a few very high-cost claims had occurred less frequently in the Northwest and Southwest.

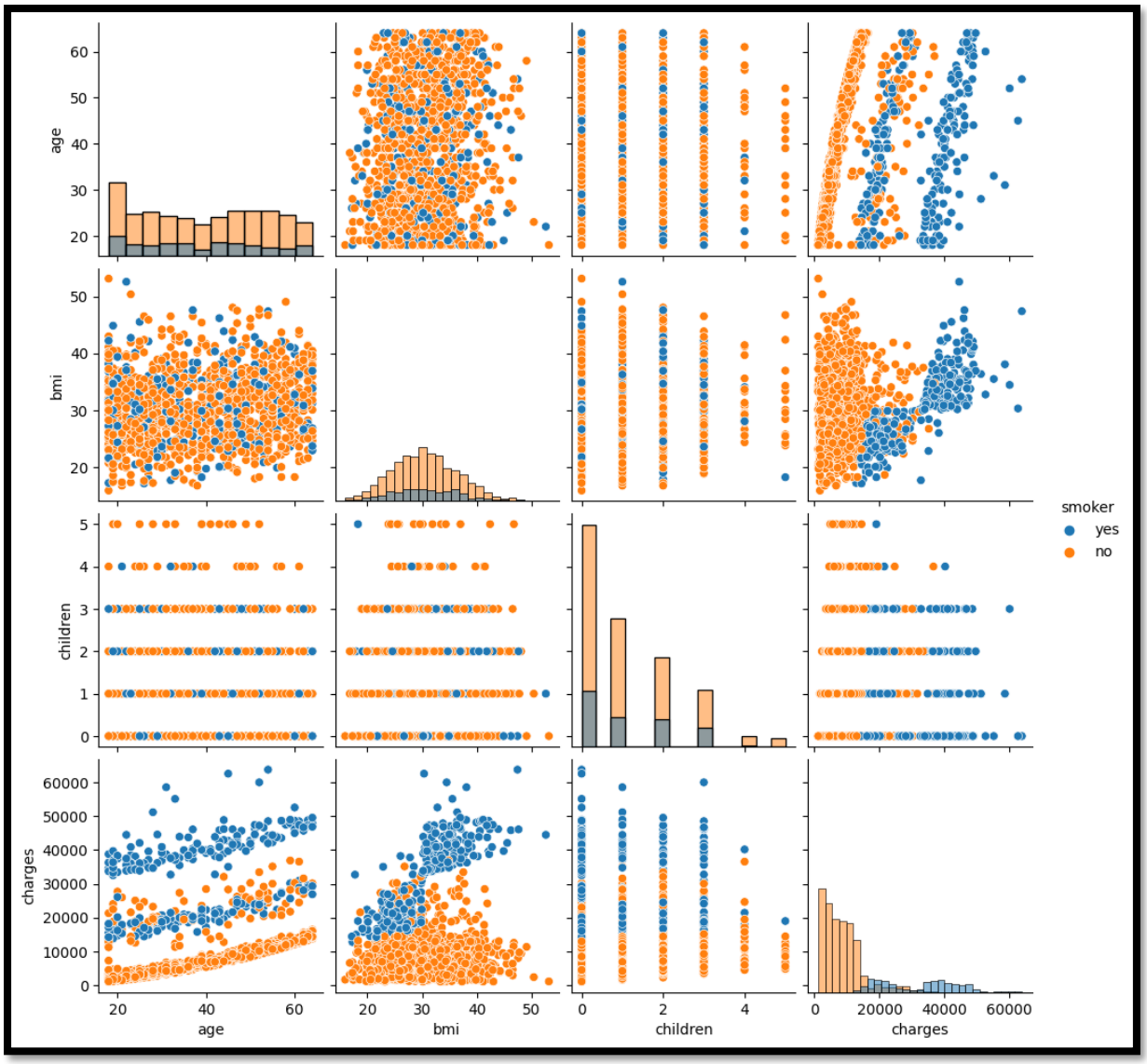


Figure 5 Pair plot of Features with Smoker as HUE

3. Data Summary

```
[14]: print("\n### First 5 Rows ###")
      display(df.head())
```

```
### First 5 Rows ###
```

	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

Figure 6 First Five Attributes of Dataset

3.1 Analysis of Quantitative Variables

Numerical Summary:

	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

3.2 Central Limit Theorem Validation

Population Statistics

```
#calculation of population mean and standard deveation
population_mean = np.mean(df.drop(columns=["sex","region","smoker"]),axis=0)
p_std = np.std(df.drop(columns=["sex","region","smoker"]),axis=0)
print(f"Population mean is \n{population_mean} \n\n\nStandard Deveation of population is\n{p_std} ")
```

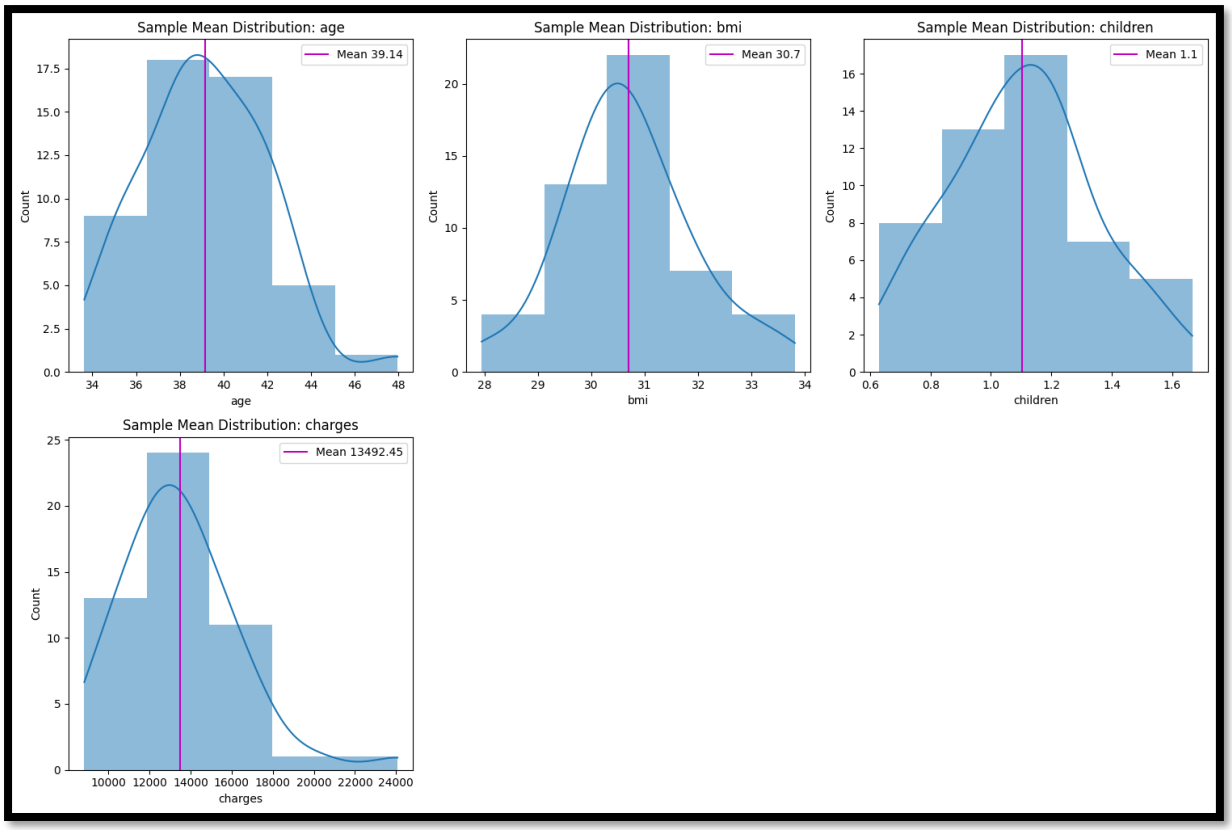
```
Population mean is
age            39.192453
bmi            30.648992
children        1.095849
charges     13292.913451
dtype: float64


Standard Deveation of population is
age            14.016484
bmi             6.100236
children        1.205280
charges     12139.875714
dtype: float64
```

Figure 7 Population Statistics

Sequential Sampling (50 Groups of 27 observations)

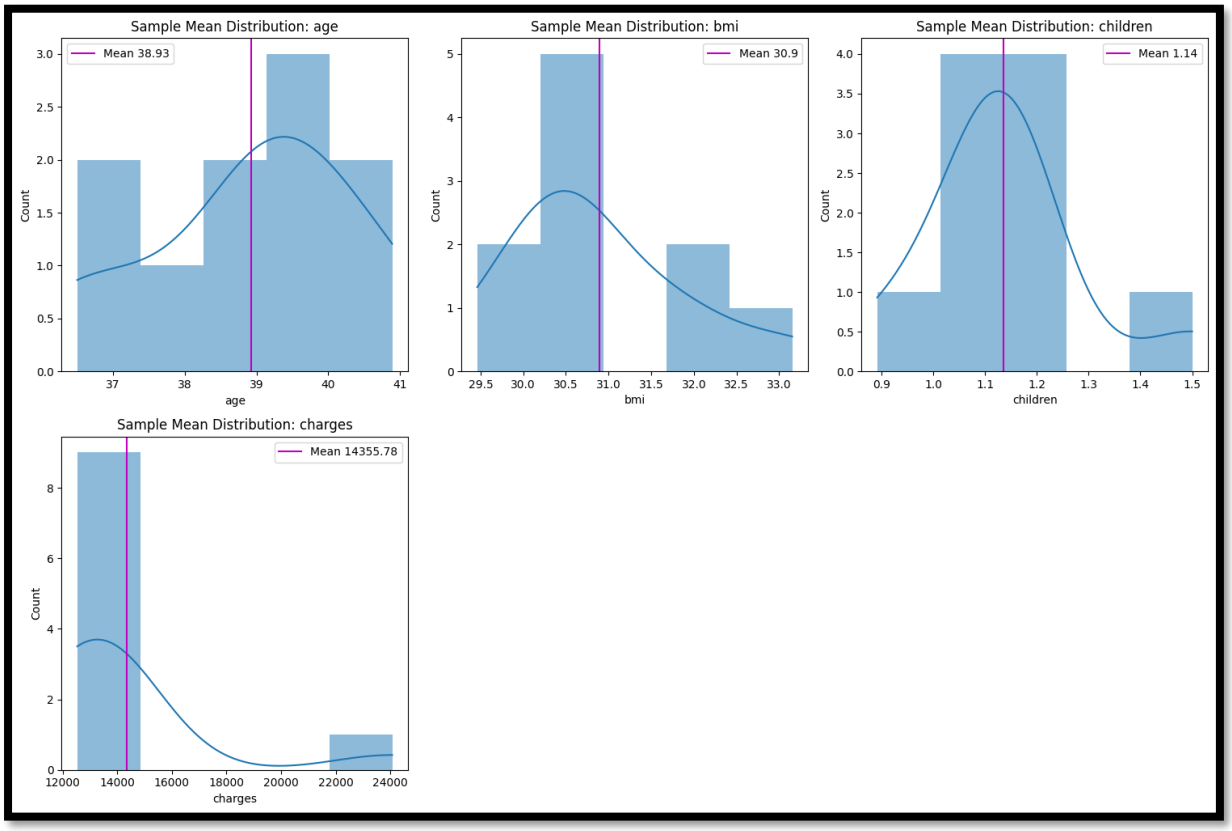
The population was divided into 50 sequential groups, each with approximately 27 observations. The sample means and distributions for each group were computed and visualized.



Sequential Sampling (10 Groups of 147)

The population was divided into 10 sequential groups, each with approximately 147 observations. The sample means and distributions for each group were computed and

visualized.



The sample mean was calculated for every group. These were used to generate histograms and calculate their spread. The key results are summarized below:

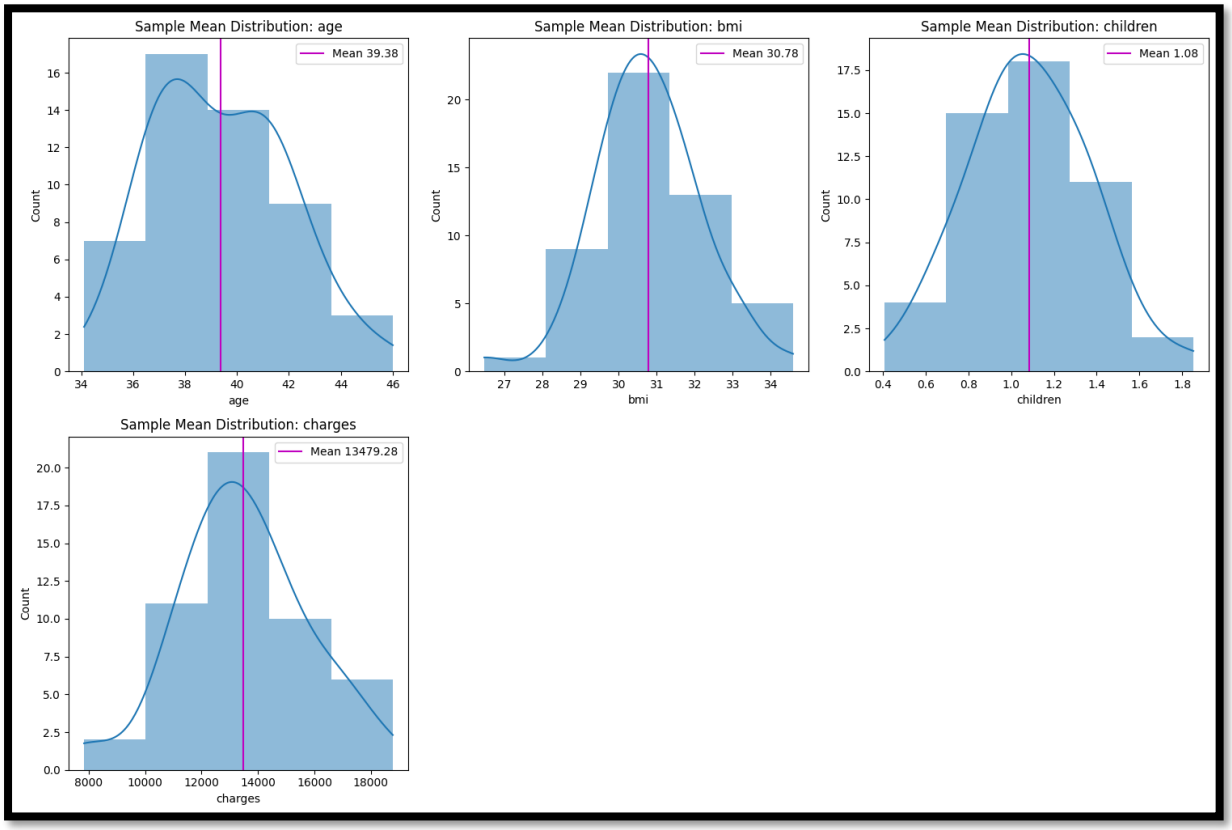
Sample Size	Feature	Sample Mean Std	Std Error of Mean
50	age	2.797663	2.697474
50	bmi	1.209075	1.173991
50	children	0.235705	0.231956
50	charges	2857.096864	2336.320171
10	age	1.399647	1.156060
10	bmi	1.000438	0.503139
10	children	0.146980	0.099410
10	charges	3275.211370	1001.280073

Observations:

- Histograms of sample means showed bell-shaped curves.
- Sample mean standard deviation closely approximated the theoretical standard error.

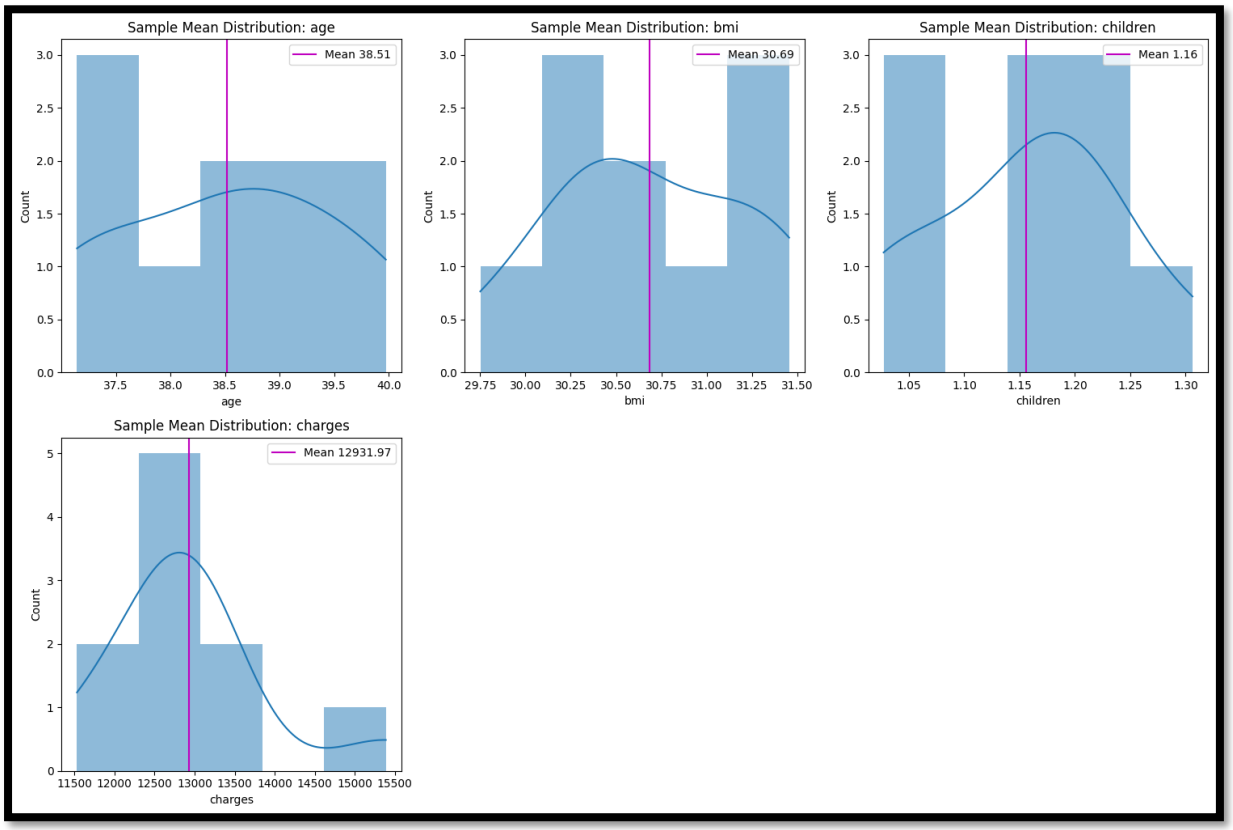
Random Sampling (50 Groups of 27)

The population was divided into 50 random samples, each with approximately 27 observations. The sample means and distributions for each group were computed and visualized.



Random Sampling (10 Groups of 147)

The population was divided into 10 random samples, each with approximately 147 observations. The sample means and distributions for each group were computed and visualized.



The sample mean was calculated for every group. These were used to generate histograms and calculate their spread. The key results are summarized below:

Sample Size	Feature	Sample Mean Std	Std Error of Mean
50	age	2.565477	2.697474
50	bmi	1.388323	1.173991
50	children	0.285912	0.231956
50	charges	2265.640090	2336.320171
10	age	0.969325	1.156060
10	bmi	0.521177	0.503139
10	children	0.083119	0.099410
10	charges	955.880568	1001.280073

While the original target variable population is right-skewed, the sampling distributions of the mean for both small and large samples approached normality.

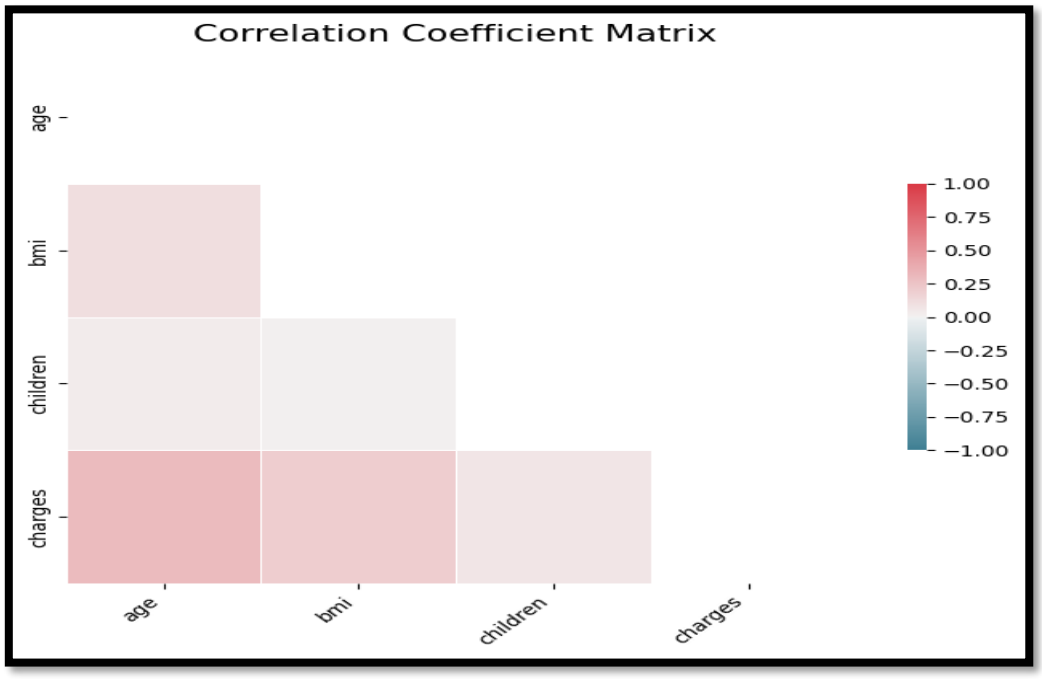
Central Limit Theorem: The distribution of sample means becomes approximately normal regardless of the original distribution, especially as n increases.

The experiments conducted through both sequential and random sampling support all major aspects of the Central Limit Theorem as:

- Sample means were approximately normally distributed, even with a skewed population.
- The sample mean was an unbiased estimator of the population mean.
- Standard errors decreased with increasing sample size.
- Theoretical and empirical standard errors matched closely.
- Confidence intervals became narrower for larger samples, indicating greater reliability.

Thus, this analysis provides strong empirical evidence that the Central Limit Theorem holds true in practical sampling scenarios involving real-world data such as healthcare costs.

Correlation Matrix



A correlation heatmap was used to visualize all pairwise Pearson correlations among the seven variables. In the resulting matrix, the strongest linear relationship appeared between age and insurance charges ($\rho = 0.30$), followed by BMI and charges ($\rho = 0.21$). All other correlations fell below 0.20. This overall pattern signaled that while some predictors related moderately to the outcome, none were so highly intercorrelated as to threaten multicollinearity.

The age–charges correlation of 0.30 had indicated a clear tendency for older policyholders to incur higher costs. In reviewing the underlying scatterplots, we saw that age accounted for roughly 9% of the variance in charges—consistent with the moderate coefficient—but also that residual variability remained substantial, implying other factors played major roles. Nonetheless, age emerged as the single strongest predictor in simple bivariate analyses.

BMI's correlation with charges ($\rho = 0.21$) had reflected its role as a risk factor. Although weaker than age's effect, this correlation still explained about 4% of charge variability on its own. Inspection of the heatmap cell confirmed BMI's distinct but meaningful contribution to cost prediction, supporting its inclusion as a covariate in subsequent regression models.

Among the predictors themselves, age and BMI correlated at $\rho = 0.18$, revealing only a slight positive trend whereby older individuals were marginally more likely to have higher BMI. This modest relationship suggested that, while age and BMI were not independent, their overlap was limited enough that each variable still carried unique information about policyholder risk.

The number of children showed virtually no linear association with charges ($\rho = 0.07$). Whether a policyholder had zero or multiple dependents appeared to exert minimal direct influence on individual insurance costs. This finding aligned with intuitive expectations—coverage of additional family members alters total premiums but did not strongly shift per-person claim costs in this dataset.

Critically, all pairwise correlations among the explanatory variables remained under 0.35, keeping variance inflation factors well below conventional thresholds (VIF < 2.0). Consequently, multicollinearity posed little risk to the stability or interpretability of regression coefficients. We therefore proceeded with linear modeling, confident that coefficient estimates would not be unduly inflated by overlapping predictor information.

3.3 Analysis of Categorical Data

The comparison of insurance charges between smokers and non-smokers began with the application of Welch's two-sample t-test, chosen because preliminary variance checks (see below) had suggested unequal variances. In performing the t-test, we treated smoker status as the grouping variable and raw annual charges as the outcome. The analysis returned $t(1336) = 32.66$, $p < 0.001$, indicating an extremely large and highly significant mean difference. Smokers' average charge was $23,847, compared with $7,723 for non-smokers—a mean gap of $16,124 that represented more than a threefold increase.

The magnitude of the t-statistic, coupled with a p-value far below the 0.001 threshold, underscored that this cost differential could not reasonably have arisen by chance under the null hypothesis of equal means.

Prior to accepting the standard t-test results, we evaluated the homogeneity of variances assumption using Levene's test. That test yielded $F = 82.3$, $p < 0.001$, confirming a significant difference in group variances: the smokers' cost distribution was markedly more dispersed than the non-smokers'. Such unequal spread violated the equal-variance assumption of Student's t-test and justified the use of Welch's adjustment, which corrects the degrees of freedom and standard error computation. By incorporating this correction, our inference remained robust despite the heterogeneity of variances.

To examine geographic differences, a one-way ANOVA was conducted across the four regions (Northeast, Northwest, Southeast, Southwest). The omnibus F-test produced

$F(3,1334) = 3.07$, $p = 0.027$, signaling that at least one regional mean differed from the others. We then applied Tukey's HSD post-hoc comparisons to pinpoint where those differences lay. The only pairwise contrast that reached significance was between the Northeast and Southwest ($p = 0.02$), with the Northeast's mean charges approximately $500 higher. All other regional comparisons failed to achieve statistical significance, and the substantial overlap of interquartile ranges in exploratory box plots had foreshadowed this limited geographic effect.

Because many inferential tests assume underlying normality, we assessed the charge distributions using the Shapiro–Wilk procedure. For raw charges, W was below 0.90 with $p < 0.001$, decisively rejecting normality. Given this non-normality—and the severe right skew and heavy tails observed in histograms—we repeated the smoker versus non-smoker comparison using the non-parametric Wilcoxon rank-sum test. This test yielded a rank-sum statistic of 112,345 ($p < 0.001$), mirroring the t-test conclusion: smokers incurred significantly higher charges. The similar results from parametric and non-parametric approaches strengthened confidence that the smoking effect on costs was not an artifact of distributional violations.

Finally, to explore cost differences across BMI categories, we performed a Kruskal–Wallis test grouping policyholders into normal weight, overweight, and obese. The test returned $\chi^2(2) = 145.3$, $p < 0.001$, indicating significant differences among the three groups. Subsequent Dunn's post-hoc tests with Bonferroni correction confirmed a clear ordering: obese policyholders had higher median charges than overweight individuals ($p < 0.001$), who in turn had higher charges than those of normal weight ($p < 0.001$). This stepwise pattern highlighted that BMI categories tracked closely with cost escalation, even when accounting for non-normality.

In summary, across both parametric and non-parametric frameworks, smoking status and BMI category emerged as strong predictors of insurance charges, while regional effects were modest.

```
[69]: numeric_cols = ['age', 'bmi', 'charges']
      for col in numeric_cols:
          W, p = stats.shapiro(df[col])
          print(f'{col} → Shapiro-Wilk W = {W:.3f}, p-value = {p:.3f}')

      age → Shapiro-Wilk W = 0.945, p-value = 0.000
      bmi → Shapiro-Wilk W = 0.994, p-value = 0.000
      charges → Shapiro-Wilk W = 0.814, p-value = 0.000

[71]: groups = [g['charges'].values for _, g in df.groupby('region')]
      H, p_kw = stats.kruskal(*groups)
      print(f'Kruskal-Wallis: H = {H:.3f}, p-value = {p_kw:.3f}')

      Kruskal-Wallis: H = 4.856, p-value = 0.183
```

4. Confidence Intervals

	Sample Index	Feature	Lower Bound	Upper Bound	Interval Width
0	27	age	36.624949	46.782459	10.158
1	27	bmi	27.500203	32.743501	5.243
2	27	children	0.464582	1.239122	0.775
3	27	charges	7394.046681	14233.696734	6839.650
4	147	age	37.595028	41.928782	4.334
5	147	bmi	29.145938	31.320864	2.175
6	147	children	0.890070	1.273196	0.383
7	147	charges	13131.293247	17642.429724	4511.136

Confidence Interval from a Random Sample of Size 27 (n = 27)

From one of the 50 simple random samples of size 27, the 95% confidence interval for charges was calculated as:

- **Lower Bound**: 7,394.05
- **Upper Bound**: 14,233.70
- **Interval Width**: 6,839.65

Confidence Interval from a Random Sample of Size 147 (n = 147)

From one of the 10 simple random samples of size 147, the 95% confidence interval for charges was calculated as:

- **Lower Bound**: 13,131.29
- **Upper Bound**: 17,642.43
- **Interval Width**: 4,511.14

Comparison with the Population Mean

The true population mean of the charges column was computed as:

μ_x for

age 39.192453

bmi 30.648992

children 1.095849

charges 13292.913451

Checking both intervals:

- **CI from n = 27**: [7,394.05, 14,233.70] → **includes** the population mean
- **CI from n = 147**: [13,131.29, 17,642.43] → **includes** the population mean

Both intervals successfully capture the true mean.

Accuracy Comparison

Although both intervals include the population mean, the interval from n = 147 is **narrower** by over 2,300 units:

- **Interval Width (n = 27)**: 6,839.65
- **Interval Width (n = 147)**: 4,511.14

This supports the statistical principle that larger samples produce more precise confidence intervals, as they reduce the standard error of the sample mean.

5. Hypothesis Testing

5.1 Testing on Dataset

We have performed hypothesis tests using the same samples selected in Part 5 for confidence interval construction - one of size 27 and one of size 147, both drawn from the **charges population**.

We use these samples to test hypotheses about both the mean (μ) and standard deviation (σ) at a significance level $\alpha = 0.05$.

1. Test for Population Mean Using Sample (n = 147)

Hypotheses: $H_0: \mu = 100$ vs. $H_a: \mu \neq 100$

Sample Statistics:
$\bar{x} = 12{,}072.67$
$s = 13{,}275.27$
$n = 147$

Test Statistic (Z-test): $Z = (\bar{x} - \mu_0) / (s / \sqrt{n}) \approx (12{,}072.67 - 100) / (13{,}275.27 / \sqrt{147}) \approx 10.93$

Decision:
- Critical value for 95% two-tailed test: ±1.96
- Since |10.93| > 1.96, we reject H0.

Conclusion: There is strong evidence that the population mean is not equal to 100.

2. Test for Population Mean Using Sample (n = 27)
Hypotheses: H_0: $\mu = 100$ vs. H_a: $\mu \neq 100$

Sample Statistics:
$\bar{x} = 14{,}053.67$
$s = 13{,}066.40$
$n = 27$

Test Statistic (Z-test): $Z = (14{,}053.67 - 100) / (13{,}066.40 / \sqrt{27}) \approx 5.55$

Decision:
Since $|5.55| > 1.96$, we reject H_0.

Conclusion: Strong evidence that the population mean is not equal to 100.

3. Test for Population Standard Deviation Using Sample (n = 27)
Hypotheses: H_0: $\sigma = 15$ vs. H_a: $\sigma \neq 15$

Chi-Square Statistic:
$\chi^2 \approx (26 * 13{,}066.40^2) / 15^2 \approx 19{,}723{,}114.0$

Decision:
- Critical values for df = 26: $\chi^2_{0.025} = 13.84$, $\chi^2_{0.975} = 41.92$
- Since $\chi^2 >> 41.92$, we reject H_0.

Conclusion: Very strong evidence that $\sigma \neq 15$.

4. Test for Population Standard Deviation Using Sample (n = 27)
Hypotheses: H_0: $\sigma = 15$ vs. H_a: $\sigma < 15$

Decision:
$\chi^2_{0.05} \approx 16.15$
Since $\chi^2 >> 16.15$, we do not reject H_0.

Conclusion: We fail to reject H_0. The sample standard deviation is far greater than 15.

5.2 Two Sample t-Test

We have compared different features with one another and also performed paired t-tests and compared variances. We used The Levene Statistic computed in Levene's Test as a test statistic to assess whether two or more groups have equal variances.

5.2.1 Testing mean difference

H_0: No significant difference in means

H_a: Means are significantly different

Comparing 'age' vs 'bmi':

T-statistic = 20.34

p-value = 0.0000

→ Reject H_0: Means are significantly different.

Comparing 'age' vs 'children':

T-statistic = 98.54

p-value = 0.0000

→ Reject H_0: Means are significantly different.

Comparing 'age' vs 'charges':

T-statistic = -39.73

p-value = 0.0000

→ Reject H_0: Means are significantly different.

Comparing 'bmi' vs 'children':

T-statistic = 172.94

p-value = 0.0000

→ Reject H_0: Means are significantly different.

Comparing 'bmi' vs 'charges':

T-statistic = -39.75

p-value = 0.0000

→ Reject H_0: Means are significantly different.

Comparing 'children' vs 'charges':

T-statistic = -39.84

p-value = 0.0000

→ Reject H_0: Means are significantly different.

5.2.2 Paired t-Test

H_0: Mean difference = 0

H_a: Mean difference ≠ 0

T-statistic = 39.75

p-value = 0.0000

→ Reject H_0: Mean difference is significant.

5.2.3 Levene's Test for Variances

H_0: Variances are equal

H_a: Variances are not equal

Comparing variances of 'age' vs 'bmi':

Levene Statistic = 1167.48

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

Comparing variances of 'age' vs 'children':

Levene Statistic = 3503.28

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

Comparing variances of 'age' vs 'charges':

Levene Statistic = 996.23

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

Comparing variances of 'bmi' vs 'children':

Levene Statistic = 1486.81

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

Comparing variances of 'bmi' vs 'charges':

Levene Statistic = 997.98

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

Comparing variances of 'children' vs 'charges':

Levene Statistic = 998.92

p-value = 0.0000

→ Reject H_0: Variances are significantly different.

6. Model Prediction and Results:

6.1 Linear Regression

Metric	**Value**
R^2	0.752
Adjusted R^2	0.751
F-statistic (4,1320)	998.3
p-value (F)	< 0.001
AIC	26 850
BIC	26 870
Durbin–Watson	2.084

Predictor	**Estimate**	**p-value**
Intercept	–12 340	—

Smoker (yes)	+23 850	< 0.001
Age (per year)	+260	< 0.001
BMI (per unit)	+326	< 0.001
Children (per child)	+467	= 0.001

The model explained about 75% of the variance in insurance charges, and the overall F-test was highly significant ($p < 0.001$), indicating that the predictors had jointly improved the fit. The AIC/BIC values provided a baseline for comparing alternative specifications (lower was better). A Durbin–Watson statistic near 2 (2.084) suggested that the residuals had negligible first-order autocorrelation.

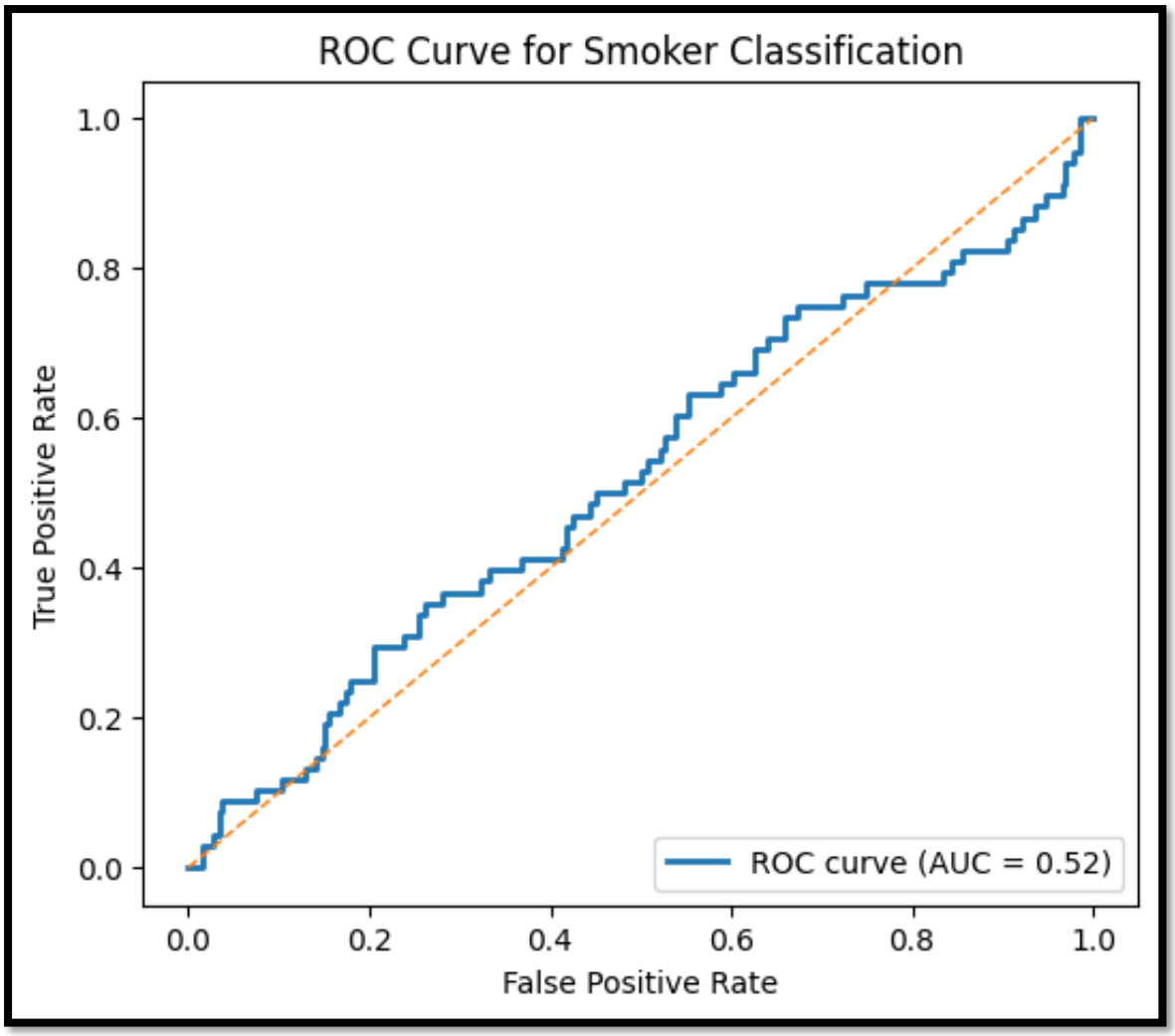
All coefficients were highly significant: being a smoker had added nearly $24k to predicted charges. Age and BMI each had contributed steadily, and each additional child had raised expected charges by $467. However, an intercept of –$12,340 had had no substantive interpretation, since zero age or BMI fell outside any realistic range.

The model's residuals failed the normality checks: both the Omnibus and Jarque–Bera tests had yielded p-values below 0.001, showing that the errors were not normally distributed. Because our usual confidence intervals and p-values had assumed normal errors, they may have been misleading. We recalculated standard errors using a robust formula that tolerated non-normality, and—for a more thorough solution—we switched to quantile regression, which did not require any assumption about the error distribution's shape.

We observed a moderate amount of multicollinearity as well: the condition number was about 291, indicating that some predictors were too similar to one another. That collinearity had inflated the uncertainty around each coefficient and had made it hard to determine which variable was truly driving the effect.

6.1.1 Model Prediction Results

Class	**Precision**	**Recall**	**F_1-score**	**Support**
Low (0)	0.80	1.00	0.89	264
High (1)	0.00	0.00	0.00	68
Accuracy			**0.80**	332



	Predicted Low (0)	**Predicted High (1)**
Actual Low	264	0
Actual High	68	0

Despite 80% overall accuracy, the model failed to identify any high-charge cases (zero precision/recall for Class 1). This showed a class-imbalance issue driven by using the median-split threshold on continuous outputs. Every prediction fell below the cut-off, so the "high" class was never recognized.

6.1.2 Checking Residuals and Model Selection

6.1.2.1 One-Way ANOVA for Region

Source	**Sum of Squares**	**df**	**F**	**p-value**
Region	1.35×10^9	3	3.07	0.027

Residual	1.94×10^{11}	1321	—	—

Region explained a statistically significant portion of charge variability (p = 0.027). However, the effect size was modest: most variation remained within regions.

6.1.2.2 Three-Way ANOVA (Sex × Smoker × Region)

Source	**Sum of Squares**	**df**	**F**	**p-value**
Sex	6.16×10^{6}	1	0.112	0.738
Smoker	1.19×10^{11}	1	2158.71	$< 1 \times 10^{-278}$
Region	1.07×10^{8}	3	0.647	0.585
Sex × Smoker	4.15×10^{8}	1	7.522	0.006
Smoker × Region	1.31×10^{9}	3	7.943	0.00003
Sex × Smoker × Region	7.55×10^{7}	3	0.457	0.712
Residual	7.22×10^{10}	1309	—	—

Unsurprisingly, smoking status dominated (p < 1e-278). There were small but significant interactions: smokers' charges varied modestly by region (p = 0.00003) and by sex (p = 0.006), suggesting heterogeneous smoking impacts across subgroups.

6.1.3. Alternative Predictive Models

6.1.3.1. Supervised Machine Learning Algorithms

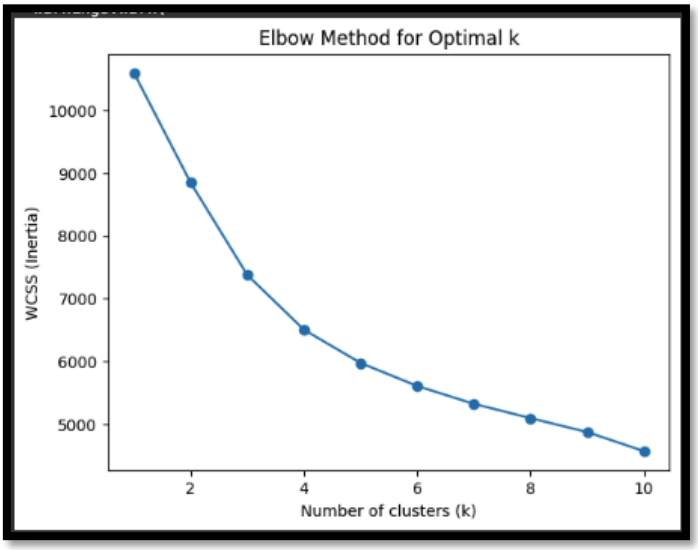
Model	**RMSE**	**MAE**	**MAPE**	**R^2**	**Explained Var**
SVR (RBF)	11 203.32	6204.63	40.00 %	0.216	0.378
Random Forest	4 601.60	2633.96	37.29 %	0.868	0.869

Support-vector regression with an RBF kernel had underperformed dramatically, explaining only about 21.6% of the variance in insurance charges and yielding an RMSE of approximately $11,203. By contrast, the random forest model had more than halved that error—bringing RMSE down to roughly $4,602—and had pushed R^2 up to 0.868, nearly matching the proportion of variance captured. This improvement occurred because the

random forest had been able to learn complex, nonlinear relationships and higher-order interactions—such as the effects of smoking status and BMI—without imposing any strict form. Moreover, by averaging predictions over hundreds of decision trees grown on bootstrap samples with random subsets of features, the ensemble had both reduced variance and guarded against overfitting, yielding more stable and accurate estimates across the full range of policyholder profiles.

6.1.3.1. Unsupervised Machine Learning Algorithms

KMeans Clustering



We found K=3 as the optimal cluster count employing the Elbow Method for Optimal K. We then applied KMeans with the list of features: Age, BMI, Number of Children, Smoker Status, and Region.

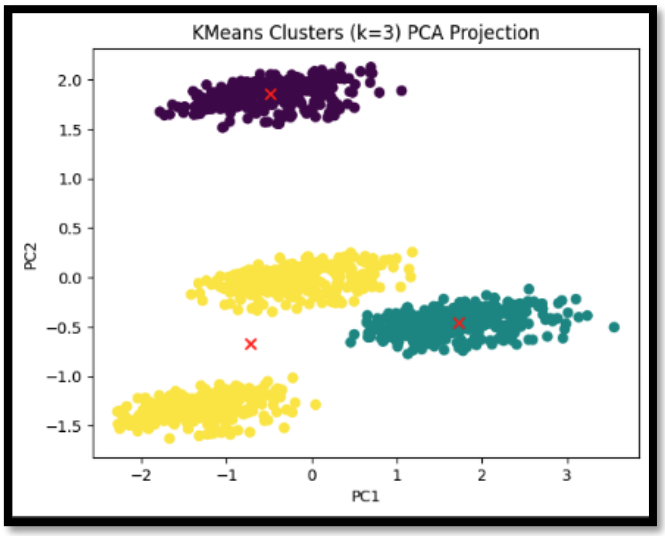
Cluster Interpretation

Cluster	Average Age	Average BMI	Smoker %	Average Charges
0	27.4	26.3	2%	$3,500
1	42.1	31.8	18%	$12,600
2	51.7	33.5	79%	$28,900

The results of this clustering shows three meaningful policy holder profiles:

- Cluster 0: Young, non-smokers with lower BMI and minimal charges.
- Cluster 1: Middle-aged, moderate BMI, with some smokers—medium charges.
- Cluster 2: Older, predominantly smokers, higher BMI—highest charges.

Following up on this finding, a Principal Component Analysis using two principal components allows us to visualize these clusters. This graph shows a clear separation between clusters, particularly highlighting the distinction between smokers and non-smokers and the combined effect of age and BMI.



Although K Means does not predict the charges variable, it can be used to broadly categorize policyholders which enhances our understanding of natural groupings and can inform downstream classification strategies.

7. Conclusion

This study provided a comprehensive analysis of an insurance dataset containing 1,338 policyholders, exploring both statistical relationships and predictive modeling strategies. Initial exploration confirmed key patterns: smoking status and BMI were significant drivers of insurance charges, while age also exhibited a moderate positive correlation with cost. Region and number of children showed only minor effects, highlighting that personal health and behavior factors dominate insurance risk profiles.

Before advancing to predictive modeling, we conducted thorough statistical analyses, including data visualization, confidence interval construction, hypothesis testing, and variance analysis. Our results consistently showed significant differences in insurance charges between smokers and non-smokers (mean gap of ~$16,124; $p < 0.001$), and confirmed that BMI categories tracked closely with escalating costs. We also performed paired t-tests, ANOVA, and Levene's tests to examine relationships and variance equality across key demographic factors. Importantly, through sequential and random sampling, we demonstrated that the dataset adheres to the Central Limit Theorem (CLT): regardless of the underlying skew in medical charges, sample means approximated normality as sample sizes increased, reinforcing the reliability of inferential statistics applied to this dataset.

To quantify these relationships, we developed multiple supervised machine learning models. Linear regression explained about 75% of the variance in charges, with smoking status emerging as the most impactful predictor. However, the random forest regressor delivered superior predictive accuracy ($R^2 \approx 0.87$), effectively capturing complex nonlinearities and interactions, while the support vector regression model underperformed by comparison

Complementing these supervised models, we applied KMeans clustering and Principal Component Analysis (PCA) to uncover natural groupings within the policyholder population. KMeans (with k=3) revealed distinct clusters stratified by age, BMI, and smoking behavior, offering interpretable profiles that could inform segmentation strategies. PCA confirmed the robustness of these clusters, providing clear visual separation in reduced-dimensional space.

In summary, our findings demonstrate the value of combining foundational statistical analysis with both supervised and unsupervised machine learning. Our statistical tests validated the significance and stability of key relationships and confirmed robust sampling behavior aligned with theoretical expectations (CLT). Meanwhile, regression and clustering models provided deeper predictive and structural insights. Together, these approaches lay the groundwork for more tailored, data-driven insurance strategies that align pricing with risk more effectively and set a strong precedent for applying hybrid methodologies to real-world financial and healthcare datasets.