Final Report: Predicting Medical Insurance Charges for a South African Medical Aid Scheme

Introduction

This report presents a predictive analysis of medical insurance charges for a South African medical aid scheme, using the Kaggle Insurance Dataset. The objective is to develop a linear regression model to predict charges based on client features (e.g., age, smoking status), providing insights for risk assessment and premium adjustments. The dataset includes 1,338 US-based records with features: `age`, `sex`, `bmi`, `children`, `smoker`, `region`, and `charges`. While US-based, it serves as a proof of concept, with a recommendation to use South African data in future analyses (Step 1). The analysis follows a structured process: evaluating dataset suitability (Step 1), planning the analysis (Step 2), conducting the analysis (Step 3), and evaluating the model (Step 4), as outlined below.

Exploratory Data Analysis (EDA)

Data Quality and Cleaning

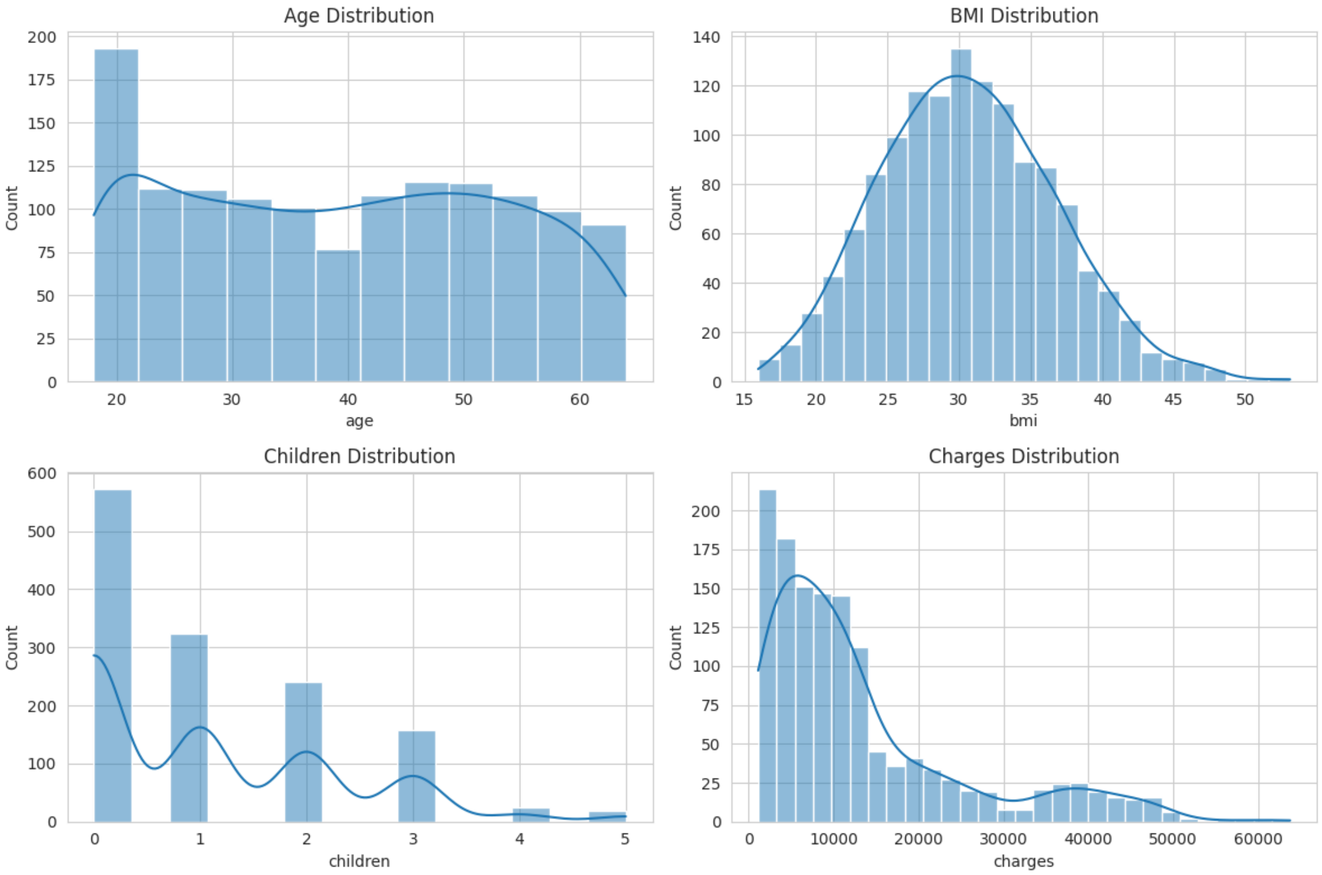
The dataset has 1,338 records and 7 features with no missing values. One duplicate was removed, resulting in 1,337 records (Step 3a). The target variable `charges` is continuous (float64, e.g., $16,884.924), meeting linear regression requirements (Step 1).

Summary Statistics

`charges` ranges from $1,121 to $63,770, with a mean of $13,279 and standard deviation of $12,110, indicating high variability. `age` averages 39.2 years, `bmi` averages 30.7, and `children` averages 1.1 (Step 3a).

Distributions

- `age` is uniformly distributed (18–64 years), with no extreme outliers.
- `bmi` is near-normal but has outliers above 45.
- `children` is right-skewed (most clients have 0–2 children).
- `charges` is right-skewed (skewness = 1.52, Step 1), with outliers above $50,000.
- Visualizations: See the histograms below:



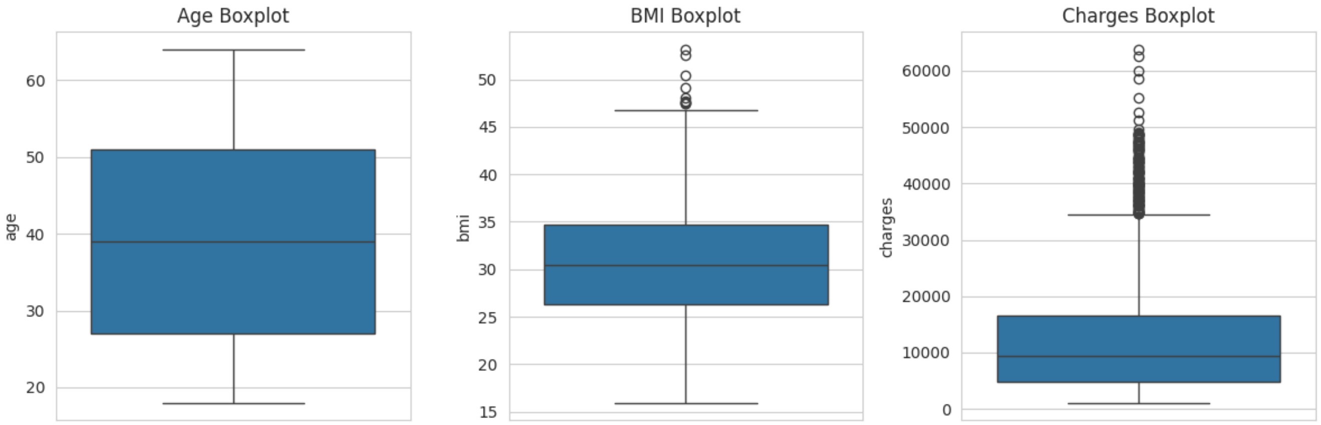
Age Distribution

BMI Distribution

Children Distribution

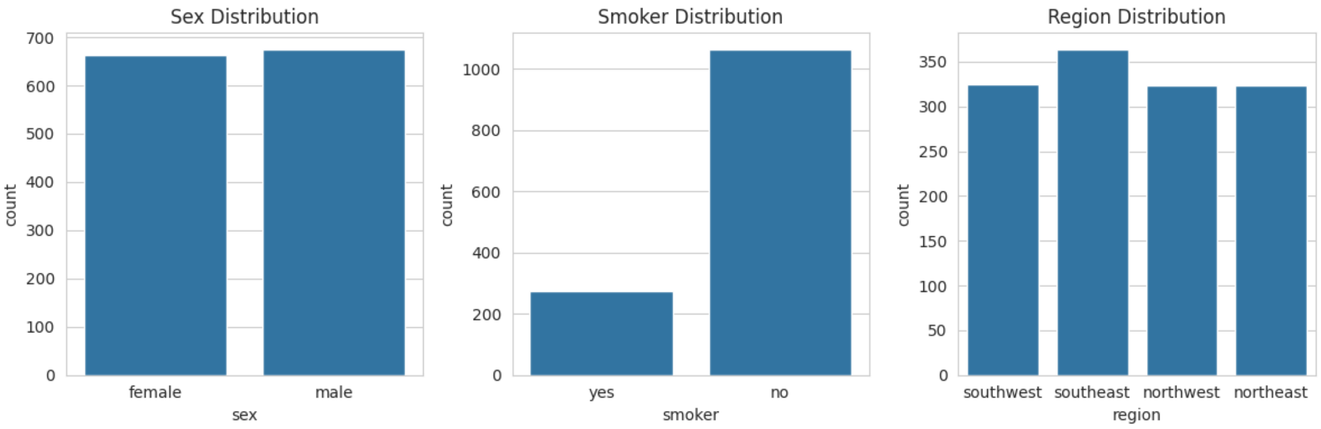
Charges Distribution

And the outliers:



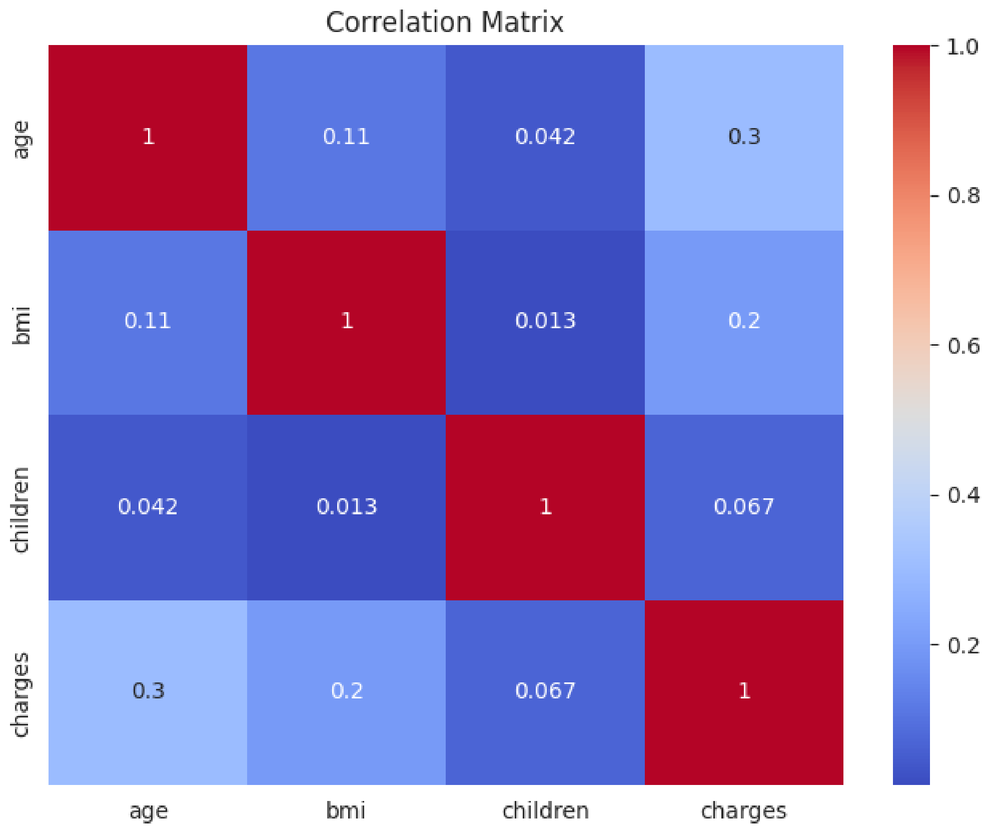
Categorical Features

- `sex` is balanced (~50% male/female).
- `smoker` is imbalanced (~20% yes), with smokers paying ~$20,000 more (see below).
- `region` is roughly equal (~25% each), with the northeast slightly higher in charges.
- Visualization: See below:

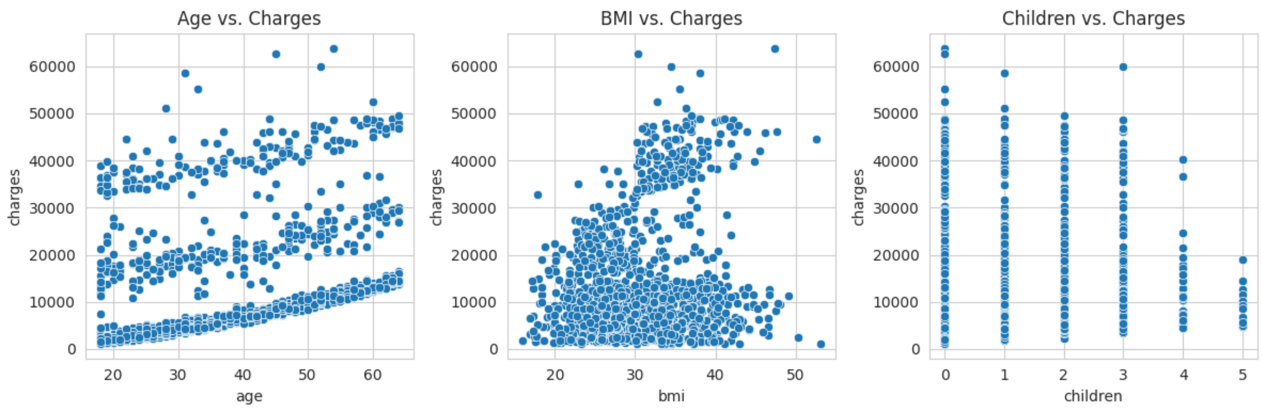


Correlations and Relationships

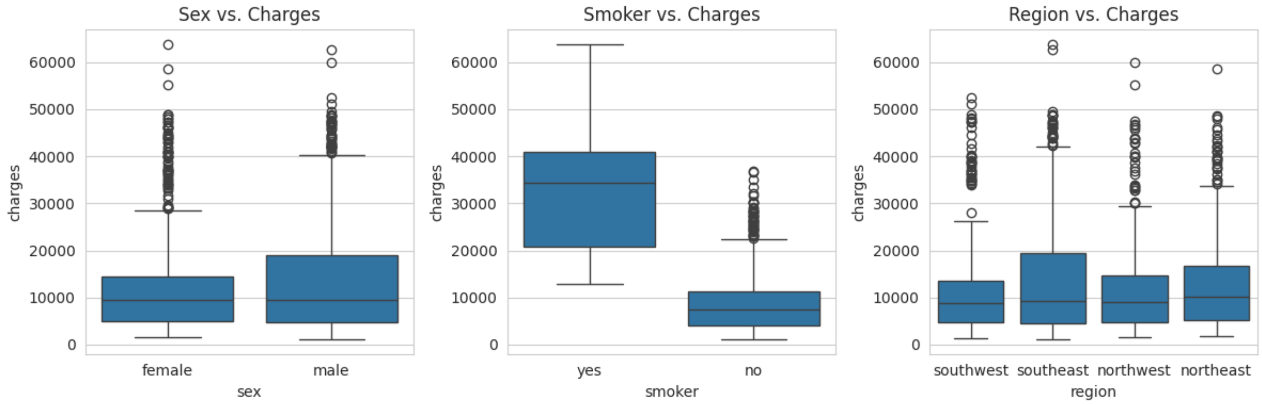
- Numerical correlations: `age` (0.3), `bmi` (0.2), and `children` (0.07) with `charges`:



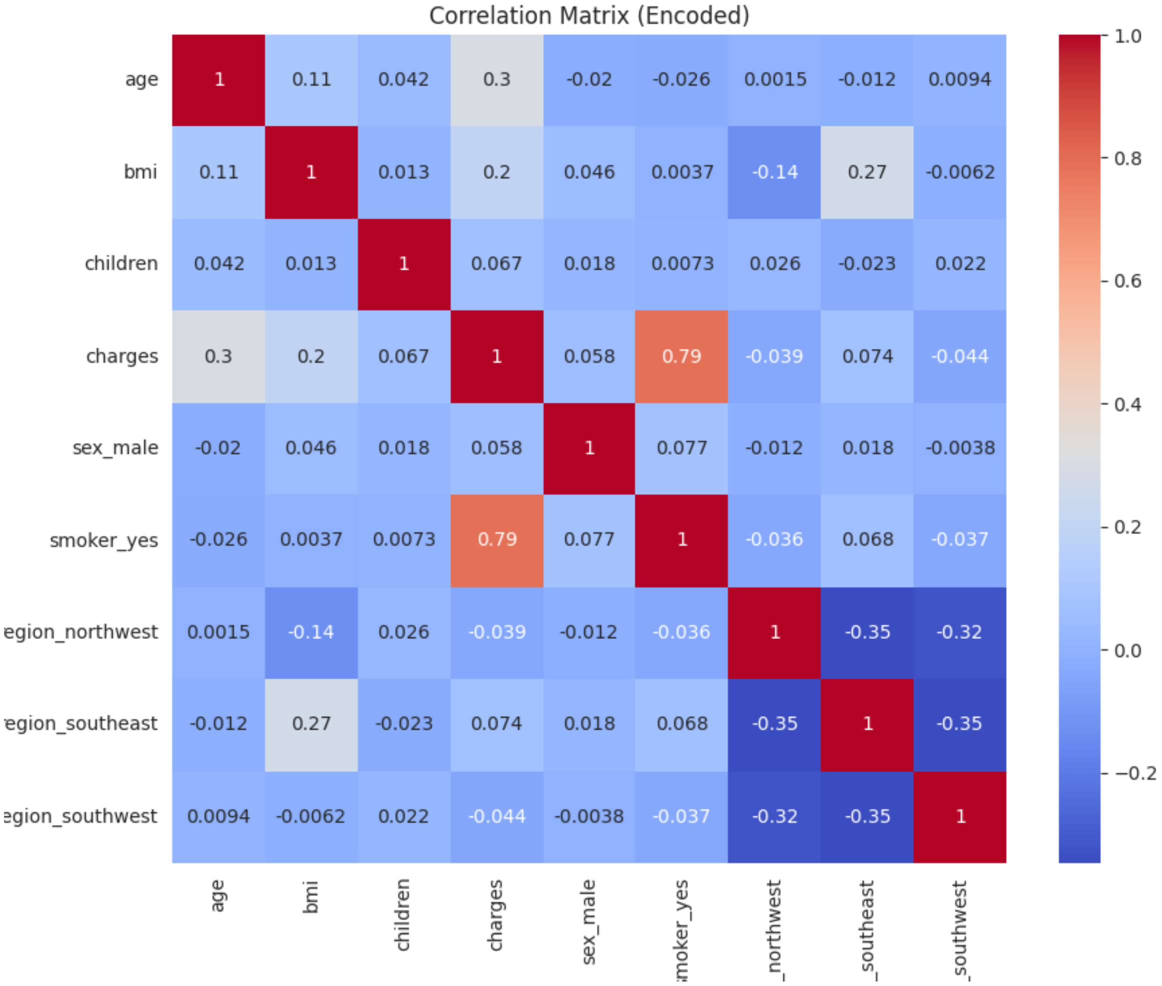
- Feature vs. `charges`:
  - `age`: Clear linear trend, a strong predictor:



  - `bmi`: Moderate trend, some outliers (see above).
  - `children`: Weak trend (see above).
  - `smoker`: Significant effect (smokers pay ~$20,000 more):



  - `sex` and `region`: Minor effects (see above).
- After encoding, `smoker_yes` (0.79), `age` (0.3), and `bmi` (0.2) strongly correlate with `charges`:



	age	bmi	children	charges	sex_male	moker_yes	_northwest	_southeast	_southwest
age	1	0.11	0.042	0.3	-0.02	-0.026	0.0015	-0.012	0.0094
bmi	0.11	1	0.013	0.2	0.046	0.0037	-0.14	0.27	-0.0062
children	0.042	0.013	1	0.067	0.018	0.0073	0.026	-0.023	0.022
charges	0.3	0.2	0.067	1	0.058	0.79	-0.039	0.074	-0.044
sex_male	-0.02	0.046	0.018	0.058	1	0.077	-0.012	0.018	-0.0038
smoker_yes	-0.026	0.0037	0.0073	0.79	0.077	1	-0.036	0.068	-0.037
egion_northwest	0.0015	-0.14	0.026	-0.039	-0.012	-0.036	1	-0.35	-0.32
egion_southeast	-0.012	0.27	-0.023	0.074	0.018	0.068	-0.35	1	-0.35
egion_southwest	0.0094	-0.0062	0.022	-0.044	-0.0038	-0.037	-0.32	-0.35	1

Implications

- `smoker`, `age`, and `bmi` are likely key predictors, aligning with Step 1's findings.
- `charges` skewness (1.52) suggests log-transformation (Brownlee, 2020).
- Outliers in `bmi` and `charges` may affect model performance, to be monitored in feature selection.

Feature Selection

Encoding

Categorical variables (`sex`, `smoker`, `region`) were encoded using one-hot encoding with `drop_first=True` to avoid multicollinearity, resulting in features: `sex_male`, `smoker_yes`, `region_northwest`, `region_southeast`, `region_southwest` (Step 3b, Learning Unit 4, LO1, Page 11).

Backward Elimination

- Removed features with p > 0.05: `sex_male` (p = 0.698), `region_northwest` (p = 0.465), `region_southwest` (p = 0.058), `region_southeast` (p = 0.136).
- Retained features: `age` (p = 0.000, coef = 257.77), `bmi` (p = 0.000, coef = 321.87), `children` (p = 0.001, coef = 472.98), `smoker_yes` (p = 0.000, coef =

23,810).

- Model fit: R-squared = 0.750, Adjusted R-squared = 0.749 (Step 3b).

Multicollinearity

- VIF scores: `age` (7.54) and `bmi` (8.06) > 5, indicating moderate multicollinearity; `children` (1.80), `smoker_yes` (1.25) < 5.
- Retained all features due to statistical significance and client relevance, noting multicollinearity as a limitation (Step 3b, Muller & Guido, 2016).

Implications

- Final features: `age`, `bmi`, `children`, `smoker_yes`.
- `smoker_yes` has the largest impact, consistent with EDA findings.
- `region` variables were removed but may be re-added for client context (South African medical aid scheme, Step 2 plan).

Model Training

Data Split

The dataset was split into 80% training (1,069 rows) and 20% testing (268 rows) with `random_state=42` for reproducibility (Step 3c).

Base Model

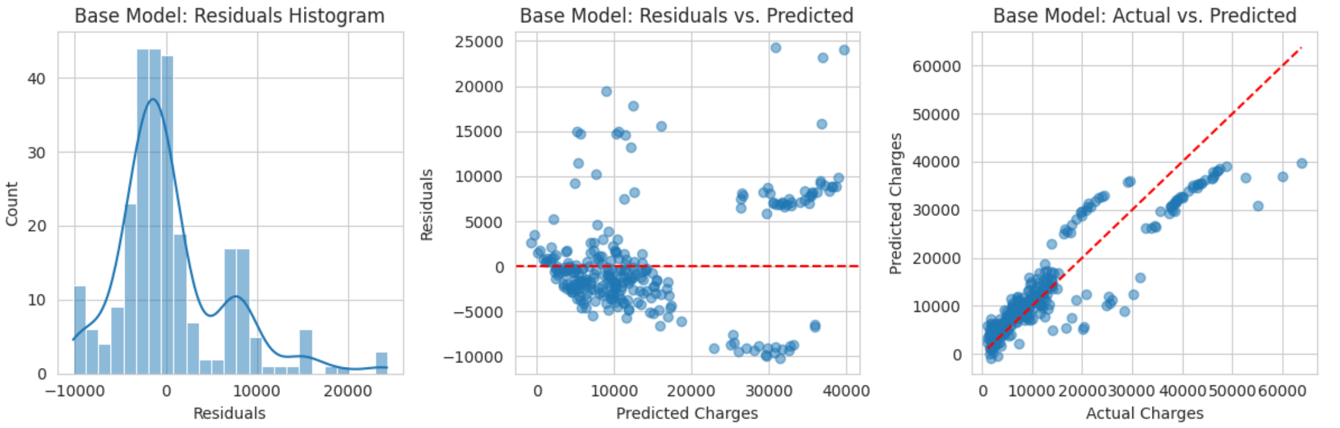
- Trained a `LinearRegression` model with default hyperparameters on features: `age`, `bmi`, `children`, `smoker_yes`.
- No hyperparameter tuning was needed, as linear regression is simple (Learning Unit 2, LO4, Page 9).

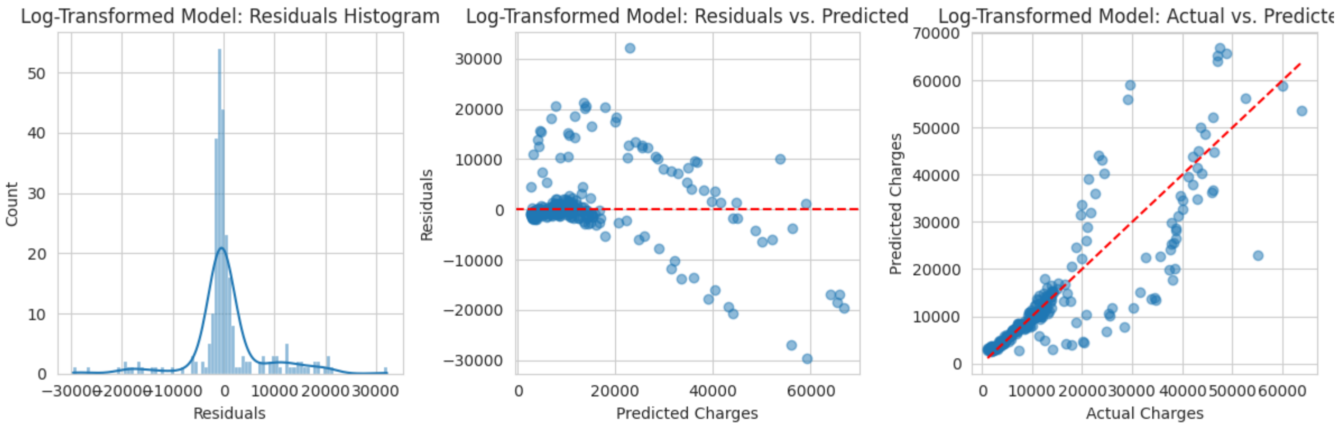
Log-Transformed Model

- Trained a second model on log(`charges`) to address skewness (1.52, Step 1), with predictions converted back to the original scale using `exp()` (Step 3c, Brownlee, 2020).

Code Snippet (from Step 3c)

```python

# Base Model Training

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42) model = LinearRegression() model.fit(X_train, y_train) y_test_pred = model.predict(X_test)

# Log-Transformed Model Training

y_log = np.log(y) X_train, X_test, y_train_log, y_test_log = train_test_split(X, y_log, test_size=0.2, random_state=42) model_log = LinearRegression() model_log.fit(X_train, y_train_log) y_test_pred_orig = np.exp(model_log.predict(X_test))

**Performance Metrics (Step 4)**

- **Base Model:**
  - MSE (Test): 33,786,990.87 - High error, reflecting variance in charges.
  - RMSE (Test): $5,812.66 - Average prediction error of ~$5,813, client-interpretable.
  - $R^2$ (Test): 0.7562 - Explains 75.62% of variance, good fit (above Step 2 threshold of 0.7).
  - Adjusted $R^2$ (Test): 0.7522 - Slightly lower, accounting for 4 features.
  - Train/test $R^2$ similar (0.7476 vs. 0.7562), no significant overfitting.
- **Log-Transformed Model (in original scale):**
  - MSE (Test): 39,568,308.52 - Higher error than base model.
  - RMSE (Test): $6,288.75 - Average error of ~$6,289, worse than base model.
  - $R^2$ (Test): 0.7150 - Explains 71.50% of variance, below base model.
  - Adjusted $R^2$ (Test): 0.7103 - Slightly lower.
  - Train/test $R^2$ similar (0.7263 vs. 0.7150), no overfitting but poorer fit.

**Residual Analysis (Step 4)**

- **Base Model:**
  - **Normality:** Residuals histogram is right-skewed (not normal), violating assumption (expected due to charges skewness = 1.52, Step 1).
  - **Homoscedasticity:** Residuals vs. predicted plot shows a funnel shape (wider spread at higher values), indicating heteroscedasticity.
- **Log-Transformed Model:**
  - **Normality:** Residuals slightly less skewed but still not normal.
  - **Homoscedasticity:** Similar funnel shape, heteroscedasticity persists.
- **Visualizations:**
```





Actual vs. Predicted

- Base model shows reasonable alignment along the diagonal, with deviations for higher charges (RMSE: $5,813).
- Log-transformed model has more deviation for higher charges (R²: 0.7150).
- See scatter plots above (`base_model_residuals.png` and `log_model_residuals.png`).

Model Coefficients (Base Model)

- `age` : $257.05 - Each additional year increases charges by ~$257.
- `bmi` : $332.89 - Each BMI unit increases charges by ~$333.
- `children` : $467.84 - Each child increases charges by ~$468.
- `smoker_yes` : $23,848.39 - Smokers pay ~$23,848 more, the largest effect (Step 4).

Discussion

Effectiveness

- The base model (R²: 0.7562, RMSE: $5,813) outperforms the log-transformed model (R²: 0.7150, RMSE: $6,289), explaining 75.62% of variance in charges. It meets the Step 2 threshold (R² > 0.7) and provides reasonable predictions (average error ~$5,813).
- Key predictors: `smoker_yes` ($23,848 increase), `age` ($257/year), `bmi` ($333/unit), and `children` ($468/child), offering actionable insights for the client.

Limitations

- **Data:** The US-based dataset may not fully reflect South African medical aid contexts (e.g., different healthcare costs, demographics). Future analyses should use SA data (Step 1).
- **Assumptions:** Both models violate normality and homoscedasticity assumptions (Step 4), likely due to outliers (charges > $50,000) and skewness (1.52). This may affect coefficient reliability.
- **Multicollinearity:** VIF for `age` (7.54) and `bmi` (8.06) > 5, indicating moderate multicollinearity, which may inflate coefficient variances (Step 3b).
- **Linearity:** The model assumes linear relationships, but residual plots suggest non-linearities (Step 4).

Recommendations for Improvement

- Remove outliers (e.g., charges > $50,000) to reduce skewness and improve residual normality (Step 3a).
- Test polynomial features (degree=2, Step 2 plan) to capture non-linear relationships (Learning Unit 2, LO2, Page 9).
- Apply stricter transformations (e.g., Box-Cox) or robust regression methods to address heteroscedasticity (Learning Unit 5, LO5, Page 12).
- Collect South African data to enhance relevance for the client.

Conclusion

The base linear regression model effectively predicts medical insurance charges (R²: 0.7562, RMSE: $5,813), identifying smoking as the largest cost driver ($23,848 increase), followed by age ($257/year) and BMI ($333/unit). For the South African medical aid scheme, these insights suggest focusing on smoking cessation programs and risk assessments for older clients or those with higher BMI. Despite a good fit, residual violations and multicollinearity indicate room for improvement through outlier removal, polynomial features, or local data collection. This proof of concept demonstrates the value of predictive modeling for healthcare cost management, with future work needed to tailor it to the South African context.

References

- Brownlee, J., 2020. *Data preparation for machine learning*. [e-book] Melbourne: Machine Learning Mastery. Available at: https://machinelearningmastery.com/data-preparation-for-machine-learning/ [Accessed 25 April 2025].
- Kaggle, 2021. Insurance dataset. Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance [Accessed 25 April 2025].
- Muller, A.C. and Guido, S., 2016. *Introduction to machine learning with Python*. Sebastopol: O'Reilly.
- Scikit-learn, 2021. User guide: linear models. Available at: https://scikit-learn.org/stable/modules/linear_models.html [Accessed 25 April 2025].
- Srivastava, T., 2019. 11 important model evaluation metrics. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ [Accessed 25 April 2025].idhya.
- Kaggle. (2021). *Insurance Dataset*. https://www.kaggle.com/datasets/mirichoi0218/insurance