Assignment MS9004

# Introduction

This report provides a comprehensive analysis of the SENIC dataset, with a focus on predicting the risk of nosocomial infections in hospitals. The study encompasses exploratory data analysis, regression modelling, diagnostic evaluations, and model refinement to enhance predictive accuracy.

## Explore Data

An exploratory analysis was conducted to examine the distribution of variables, identify potential outliers, and assess correlations among predictors. A comprehensive dataset summary and correlation analysis were performed to provide deeper insights into variable relationships.

The dataset comprises 113 observations, with no missing values across any variables. It includes eight quantitative predictors and two qualitative predictors, alongside the response variable representing infection risk.

A table with numbers and letters

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

# Stats around the data

A table with numbers and letters

AI-generated content may be incorrect.

The dataset comprises 113 observations across nine variables of quantitative predictors. The mean infection risk is 4.35, with a standard deviation of 1.34, indicating moderate variation among hospitals. The length of stay ranges from 6.7 to 19.56 days, with an average of 9.65 days. The age distribution suggests a relatively older patient population, with a mean of 53.23 years and a standard deviation of 4.46. The number of cultures performed exhibits significant variability, with a mean of 15.79 but a maximum of 60.5, suggesting that certain hospitals conduct substantially more tests than others. Similarly, the number of X-rays performed per hospital varies widely, with values ranging from 39.6 to 133.5, and an average of 81.63.

A graph with lines and dots

AI-generated content may be incorrect.The number of hospital beds, patient counts, and nursing staff also show high variability. The number of beds per hospital has an average of 252 but ranges from as few as 29 to as many as 835. The patient count varies dramatically, from 20 to 791, with an average of 191, indicating significant differences in hospital capacity. The nursing staff numbers follow a similar pattern, averaging 173.25 but peaking at 656, which may reflect differences in hospital sizes and staffing policies. Additionally, the number of facilities within each hospital ranges from 5.7 to 80, with a mean of 43.16, suggesting disparities in available resources. These variations highlight the diverse hospital characteristics and underline the need for robust predictive modelling to assess the risk of nosocomial infections accurately.

The box plot highlights variability and outliers across numerical variables. Risk, age, length, cultures, and X-rays have compact distributions, while bed, patient, and nurse counts show wide spreads with extreme values, reflecting hospital capacity differences. Facilities exhibit moderate variability, indicating resource disparities.

A pie chart with different colored circles

AI-generated content may be incorrect.A graph of gray rectangular bars

AI-generated content may be incorrect.

The visualizations provide insights into the distribution of affiliation and regions. The first pie chart shows that 85% of the data corresponds to "No" affiliation, while only 15% corresponds to "Yes," indicating a strong imbalance in affiliation responses. The second pie chart represents the distribution of regions, with the South (S) region having the highest proportion (32.7%), followed by the West (W) at 14.2%, Northeast (NE) at 24.8%, and an additional region (28.3%).

The bar chart further breaks down the affiliation distribution by region. Most regions have a higher count of "No" affiliation, with varying proportions of "Yes." The highest count appears in certain regions, while others have a lower presence of "Yes" responses. These visualizations highlight regional disparities in affiliation responses, which may be important for further analysis.

A graph of distribution curve

AI-generated content may be incorrect.

The risk distribution is approximately normal, with a slight negative skew (-0.1182) indicating a marginally longer left tail. The kurtosis (0.1218) suggests slightly heavier tails, but the deviation is minimal. Overall, the data shows near-symmetry with a distribution close to normal.

A graph of distribution of length

AI-generated content may be incorrect.

The distribution is highly right-skewed (skewness: 2.0414), indicating that most data points are concentrated on the left, with a few extreme high values extending the right tail. The high kurtosis (7.6718) suggests a leptokurtic distribution, characterized by a sharp peak and heavy tails, indicating the presence of more extreme outliers. These characteristics violate normality assumptions.

A graph of age distribution

AI-generated content may be incorrect.

The distribution is nearly symmetrical (skewness: -0.1026) with a slight leftward skew, indicating minimal asymmetry. The kurtosis (0.9672) suggests a mesokurtic distribution, closely resembling a normal distribution with neither heavy nor light tails. Overall, the data exhibits no significant deviations from normality, making it well-suited for statistical analysis that assume a normal distribution.

A graph of a number of people

AI-generated content may be incorrect.

The distribution is moderately right-skewed (skewness: 1.5887), indicating that most data points are concentrated on the left, with a few larger outliers extending the right tail. The high kurtosis (3.7406) suggests a leptokurtic distribution, characterized by a sharper peak and heavier tails, meaning extreme values are more frequent. These characteristics suggest that normality assumptions may not hold.

A graph of distribution of xray value

AI-generated content may be incorrect.

The distribution is nearly symmetrical (skewness: 0.0078), showing no significant skew in either direction. The negative kurtosis (-0.2813) indicates a platykurtic distribution with a slightly flatter peak and lighter tails, suggesting fewer extreme values. Overall, the data is close to normal, making it suitable for statistical analyses that assume normality.

A graph of a bed value

AI-generated content may be incorrect.

The distribution is moderately right-skewed (skewness: 1.3602), with most data clustered on the left and some higher outliers extending the right tail. The kurtosis (1.1728) indicates a mesokurtic to slightly platykurtic shape, meaning the peak is moderate and tails are not excessively heavy. The data is close to normal.

A graph of a patient

AI-generated content may be incorrect.

The distribution is moderately right-skewed (skewness: 1.3610), with most data clustered on the left and a few large outliers extending the right tail. The kurtosis (1.6021) suggests a slightly flattened peak with mildly lighter tails, indicating fewer extreme values than a leptokurtic distribution. The data is close to normal.

A graph of a function

AI-generated content may be incorrect.

The distribution is nearly symmetrical (skewness: 0.0732), with no significant skew in either direction. The negative kurtosis (-0.4526) indicates a platykurtic shape, meaning a slightly flatter peak and lighter tails with fewer extreme values. Overall, the data is close to normal.

A graph of a nurse

AI-generated content may be incorrect.

The distribution is moderately right-skewed (skewness: 1.3604), with most values concentrated on the left and some high-value outliers extending the right tail. The kurtosis (1.4330) indicates a mesokurtic to slightly platykurtic shape, meaning a lower peak and lighter tails compared to a normal distribution. While the data is normal, the presence of outliers may impact statistical modelling.

While infection risk has moderate variation, factors like hospital beds, staff, and patient counts differ significantly. Most data follow a normal distribution, but some have outliers and skewed values, which could impact statistical modelling. Visualizations highlight an imbalance in hospital affiliations and regional differences, with most hospitals not affiliated and varying distributions across regions.

A chart of a diagram

AI-generated content may be incorrect.

As per the pair plot, several variables, such as bed count, patient count, and nursing staff, exhibit strong positive correlations, as seen in their scatterplots with well-aligned trends. Infection risk and length of stay display moderate relationships, suggesting that prolonged hospitalization may be linked to higher infection risks. On the other hand, age and number of cultures performed show weak or no clear correlation with other variables, implying minimal influence.

A screenshot of a graph

AI-generated content may be incorrect.

The correlation analysis highlights key relationships between hospital variables. Infection risk is moderately linked to length of stay (0.53) and cultures performed (0.56), suggesting longer stays and increased testing may raise infection rates. Larger hospitals tend to have longer stayed and higher patient counts, as shown by moderate correlations (0.41–0.47). Strong correlations among hospital size indicators (above 0.90) confirm that larger facilities accommodate more patients and staff. X-ray and culture numbers also correlate with infection risk, emphasizing their role in infection detection. Notably, patient age shows little influence. Facilities also show strong correlations (above 0.75) with these variables.

# Fit the MLR model & Evaluate Model

A screenshot of a computer code

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

Affiliation\_Yes = 1 if affiliated, 0 otherwise.

Region\_NE, Region\_S, Region\_W = 1 if the hospital is in the respective region, 0 otherwise.

The analysis identifies several statistically significant predictors of risk (p < 0.05). Specifically, being in the Western region (region[T.W]) shows a positive association with risk (p = 0.009). Additionally, longer hospital stays (length, p = 0.002), a higher number of cultures (p < 0.001), and more X-rays (p = 0.010) are all associated with an increase in risk. The number of nurses (p = 0.046) also has a significant positive correlation with risk. However, several predictors do not show strong evidence of association with risk (p > 0.05), including affiliation (T.yes), other regions (NE and S), age, number of beds, number of patients, and available facilities, suggesting that these variables do not contribute significantly to explaining variations in risk within this model.

The model demonstrates a reasonable fit, with an R² value of 0.647, indicating that approximately 64.7% of the variation in the dependent variable (risk) is explained by the independent variables. The Adjusted R² of 0.587 accounts for the number of predictors, suggesting that the model still explains a substantial proportion of the variance. The Mean Squared Error (MSE) of 0.743 represents the average squared difference between actual and predicted values, where lower values indicate better model performance. Additionally, the Akaike Information Criterion (AIC) of 225.3 and the Bayesian Information Criterion (BIC) of 256.9 serve as model selection criteria, with lower values indicating a more optimal balance between model fit and complexity. F-stats p -value < 0.005.

Jarque\_bera = (1.4245062841317742, 0.4905377002540259, 0.0038368462347042076, 2.3620792291278594)

Omnibus = NormaltestResult(statistic=2.0815318436544357, pvalue=0.3531840669511357)

Durbin Watson = 2.235345429323543

Multicolinearity

Intercept 178.9280038819232

affiliation[T.yes] 2.119014783944837

region[T.NE] 2.0234824358087495

region[T.S] 1.9077985264937107

region[T.W] 1.5410726099435827

length 2.545611715793667

age 1.265624464430535

cultures 1.6188095304883192

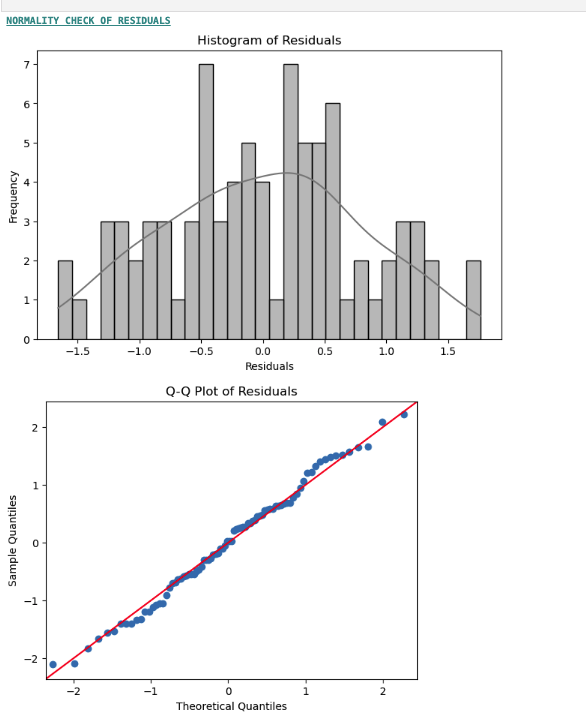
xray 1.5639853916811752

bed 47.0708848913553

patient 43.4624110789282

nurse 6.321661820417416

facilities 3.3159629523983085

 A screenshot of a graph

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

The model's assumptions and diagnostics indicate that the residuals are normally distributed, as confirmed by the Jarque-Bera (p = 0.49) and Omnibus (p = 0.35) tests. Additionally, the Durbin-Watson statistic (2.23) suggests no significant autocorrelation in the residuals, indicating that errors are independently distributed. However, multicollinearity is a concern, with severe multicollinearity detected in bed (VIF = 47.07) and patient (VIF = 43.46), while other variables exhibit acceptable VIF values.

# Improve the model PCA

This model is the preferred choice due to its improved efficiency, lower multicollinearity, and better variable selection. While the base model has a slightly higher R2 (0.647 vs. 0.629), this model achieves a more parsimonious fit with lower AIC (223.5 vs. 225.3) and BIC (247.8 vs. 256.9). By removing highly correlated and non-significant variables (bed, patient, nurse, and facilities) and adding PC1 as a significant predictor. It reduces redundancy while maintaining key explanatory power. Additionally, its lower variance inflation factors (VIF) indicate a more stable model. However, it exhibits heteroscedasticity (Breusch-Pagan p = 0.044), suggesting the need for robust standard errors.

## A screenshot of a computer AI-generated content may be incorrect.A screenshot of a computer code AI-generated content may be incorrect.

A group of blue dots

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

## Mean centring

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

The mean centre model leads slight improvement in 𝑅2 (0.642) compared to PCA (0.629), but it introduces extreme multicollinearity issues (nurse0 and facilities0 have infinite VIFs), making the model unstable. Although PCA has a slightly lower 𝑅2 it has the lowest AIC (223.5) and BIC (247.8), making it the most parsimonious.

Additionally, Model PCA removes multicollinearity-heavy variables while keeping all significant predictors intact. While mean-centring resolves some collinearity issues by centring, its condition number (4.45e+16) is excessively high, signalling serious numerical instability.

Therefore, Model PCA remains the best option, offering the optimal trade-off between predictive power, simplicity, and model stability.

## Mean-centring with PCA

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect. A graph of a graph

AI-generated content may be incorrect.

Among the models, PCA and PCA mean centred emerge as the strongest contenders, offering a balance between predictive power, multicollinearity reduction, and model efficiency. Original model has the highest R2(0.647), but it includes unnecessary variables (bed, patient, nurse, and facilities) that contribute to severe multicollinearity (VIFs above 40). Mean centring, slightly improving collinearity but introducing extreme instability (VIFs reaching infinity for nurse0 and facilities0). PCA alongside mean centring, achieves a competitive R2 (0.631) with a dramatically reduced condition number (101.97 vs. 4.45e+16 just mean centred), making it the most stable version among the transformed models. Compared to PCA alone, which also maintains a low AIC (223.5) and BIC (247.8), Mean centred with PCA performs similarly (AIC = 223.0, BIC = 247.3) but with better multicollinearity control and a lower condition number. However, Mean centred with PCA has some heteroscedasticity (Breusch-Pagan p = 0.038), which suggests the need for robust standard errors.

## Interaction term with PCA

A screenshot of a computer

AI-generated content may be incorrect.A graph with blue lines and a white background

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

Among the five models, Model interaction terms with PCA emerges as the strongest performer, achieving the highest R2(0.682) and adjusted R2 (0.629), indicating the best fit and explanatory power.

By incorporating Interaction terms with PCA between affiliation and region, it captures additional variation in risk. Moreover, it has the lowest AIC (216.4) and a reasonable BIC (248.0), making it both statistically powerful and efficient. The mean centring with PCA and PCA also perform well, with mean centring with PCA having lower multicollinearity (condition number = 101.97) and strong predictor significance, while PCA alone maintains interpretability and simplicity with the lowest AIC among the non-interaction models.

Model mean centring slightly improves upon original model but suffers from extreme multicollinearity issues (infinite VIFs), making it unstable. The original model, though achieving the second-highest R2 (0.647), retains highly collinear and unnecessary predictors (bed, patient, nurse, and facilities), reducing its reliability. While Model interaction terms with PCA is the best in terms of fit and efficiency, it does exhibit heteroscedasticity (Breusch-Pagan p = 0.035), requiring robust standard errors. If interpretability and stability are priorities, Model PCA mean centring is the best choice, while Model PCA remains a strong contender for simplicity. Ultimately the interactive model is the most comprehensive model, offering the best explanatory power and efficiency, provided heteroscedasticity is addressed.

A screenshot of a computer

AI-generated content may be incorrect. A screenshot of a computer

AI-generated content may be incorrect.

The Model interaction terms with PCA and polynomial transformation stand out as the best-performing models in terms of explanatory power and efficiency, with the highest R2 (0.682) and adjusted R2(0.629 and 0.628, respectively). Both capture additional variation in risk through interaction and nonlinear effects.

However, polynomial transform introduces a quadratic age term, which is not statistically significant (p=0.732), suggesting that age does not have a strong nonlinear relationship with risk. The interaction term model, on the other hand, reveals a significant interaction effect (affiliation × times× region S, p=0.003), making it more interpretable.

Mean centring with PCA and PCA are strong contenders for stability and interpretability. Mean centring with PCA has the lowest condition number (101.97), indicating minimal multicollinearity, while PCA alone provides a simpler model with a lower AIC (223.5 vs. 223.0 in Mean centring with PCA) and avoids transformations. However, PCA alone has a slightly lower R2 (0.629) compared to Mean centring with PCA (0.631), and Mean centring with PCA maintains better collinearity control.

Mean centring and the original model perform worse due to severe multicollinearity and inclusion of non-significant variables. Mean centring, despite centring, still suffers from infinite VIFs for some variables, making it unstable, while base model has the highest VIF values (above 40 for bed and patient) and retains unnecessary predictors, reducing its reliability.

In terms of efficiency, interaction terms and polynomial transformation have the lowest AIC values (216.4 and 216.5, respectively), making them the most optimal models. However, both exhibit heteroscedasticity (Breusch-Pagan p=0.035)for interaction terms and p=0.018 for polynomial transformation), requiring the use of robust standard errors.

If model simplicity and multicollinearity reduction are priorities, Mean centring with PCA is the best choice. If interpretability and a strong model fit are more important, interaction term is the top recommendation. Polynomial transformation provides a slightly different nonlinear perspective but does not significantly outperform interaction term in practical terms.

## Conclusion

A screenshot of a spreadsheet

AI-generated content may be incorrect.

Based on the comparative analysis of all six models, the Model interaction terms with PCA emerges as the best overall choice due to its highest adjusted R2 (0.629), lowest mean squared error (0.6685), and lowest AIC (216.4), indicating the best balance between predictive accuracy and model efficiency.

While Polynomial transformation achieves similar performance, it does not provide a significant improvement over interacting with variables and introduces non-significant polynomial terms. PCA with mean centring offers strong multicollinearity control but does not outperform interactive term in predictive accuracy. The base model, mean-centred, and PCA-only models have lower predictive power and higher AIC/BIC values, making them less favourable.

Although the interactive term exhibits heteroscedasticity (Breusch-Pagan p = 0.035), robust standard errors can mitigate this issue, ensuring reliable inference. Therefore, the interactive term is the recommended choice for its strong predictive performance, efficiency, and ability to capture meaningful interactions, making it the most robust and interpretable model for risk prediction.