

# Infection Risk Analysis: Unraveling the Key Factors

*Arpa Banik*

## **Abstract**

In the evolving field of healthcare analytics, it is crucial to understand and predict the factors that contribute to infection risks during hospital stays. This study examines a dataset of 113 selected patients analyzing variables, like the length of stay and the frequency of X ray examinations to predict the likelihood of infections. The urgency of this research is highlighted by the need to improve outcomes and reduce healthcare associated infections which're major concerns worldwide. Previous studies in this area have often failed to provide an understanding of how patient variables relate to infection risks. Many have overlooked factors. Used overly simplistic models limiting their ability to offer detailed insights. This study aims to address these limitations by utilizing regression modeling techniques such as linear regression and sequential feature selection. We acknowledge the flaws in approaches. Strive to develop a more accurate and interpretable framework for predicting infection risks. Our analysis primarily focuses on Model M1, which explores the linear relationship, between infection risks, duration of stay and frequency of X ray examinations. We have extensively tested the significance of these variables providing evidence that at least one of them significantly impacts infection risks.

Moreover, we explore Model M2, which introduces a term called "Stay2XRay" to provide an understanding of how the duration of stay and the frequency of X rays interact, with each other. Although Model M2 only explains 38% of the variation in responses it offers insights into the complex dynamics at play. This study has achieved milestones by identifying predictors that have an impact on infection risks and creating robust regression models. Furthermore, our analyses shed light on the relationships between variables empowering healthcare professionals to make informed decisions. By revealing the influence of stay duration and X ray frequency, on infection risks this study contributes to efforts aimed at improving safety and healthcare quality. As healthcare providers tackle the challenges posed by risks our findings provide insights that can guide targeted interventions and enhance patient outcomes.

Index Terms—infection risk, x-rays, regression analysis, mediation models, interaction models

## Introduction

Healthcare professionals and analysts continually grapple with the challenge of understanding and mitigating infection risks among hospitalized patients. In this study we delve into a dataset that includes information, from 113 selected patients. Our goal is to uncover the factors that influence infection risks by examining demographic and healthcare related variables with a particular focus on two variables; hospital stay duration (Stay) and X ray frequency (Xray). Against the backdrop of the healthcare landscape our analysis aims to address questions regarding the predictors of infection risk.

### Key Questions Addressed by Data Analysis:

Our primary focus in this study revolves around assessing the significance of two variables, Stay and Xray in explaining variations in infection risk. Using regression models and statistical tests we aim to answer the following questions:

1. Are hospital stay duration (Stay) and X-ray frequency (Xray) in explaining variations in infection risk?
2. Does an increase in X ray frequency while holding hospital stay constant lead to an increase in infection risk?
3. How effective are the regression models M1(without interaction) and M2(with interaction) at predicting infection risk? What are the implications of introducing interaction terms?
4. Can we estimate infection risk using models, with certain information omitted?

Here's an overview of what you can expect from the rest of this paper:

**Data Analysis and Model Insights:** Explore findings from regression models M1 and M2, examining infection risks, stay duration, and X-ray frequency. Gain insights into the significance of these variables.

**Further Analysis and Model Refinement:** Check model quality, assess tradeoffs in model simplicity, and evaluate Model M2's performance after excluding the 'Age' variable.


**Recommendations for Healthcare Providers:** Synthesize findings into actionable recommendations for healthcare providers, focusing on monitoring stay duration, leveraging X-ray insights, and enhancing facilities.

**Conclusion and Future Directions:** Summarize key insights, stress continuous improvement in patient safety, and suggest future research areas, emphasizing the evolving nature of healthcare strategies.

## Dataset Overview

The dataset consists of 12 variables and 113 observations. To conduct the analysis, we took a sectional approach and collected data from 113 individuals who represent different cases in our study. The dataset includes features like "Stay" (the number of days patients stayed in the hospital) "XRay" (the number of X rays they received) "Age," "Beds" (the number of beds in the hospital) "Census," "Nurses" (the number of nurses in the hospital) "Facilities," and "InfctRsk" (which represents infection risk).

One interesting aspect of this dataset is that it's complete meaning there are no missing observations. However, we don't have information, about outliers, variable distributions or whether there are categorical data present. To explore these aspects further we'll need to dig. With 113 entries this dataset is relatively small which might affect how generalizable our findings are. Unfortunately, the cost details regarding the dataset are not provided, whether it involves information, ethical considerations or financial resources allocated for data collection. The dataset contains the following variables and descriptive statistics are below:



	ID	Stay	Age	InfctRsk	Culture	Xray \
count	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000
mean	57.000000	9.648319	53.231858	4.354867	15.792920	81.628319
std	32.76431	1.911456	4.461607	1.340908	10.234707	19.363826
min	1.000000	6.700000	38.800000	1.300000	1.600000	39.600000
25%	29.000000	8.340000	50.900000	3.700000	8.400000	69.500000
50%	57.000000	9.420000	53.200000	4.400000	14.100000	82.300000
75%	85.000000	10.470000	56.200000	5.200000	20.300000	94.100000
max	113.000000	19.560000	65.900000	7.800000	60.500000	133.500000

	Beds	MedSchool	Region	Census	Nurses	Facilities
count	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000
mean	252.168142	1.849558	2.362832	191.371681	173.247788	43.159292
std	192.842687	0.359097	1.009437	153.759564	139.265390	15.200861
min	29.000000	1.000000	1.000000	20.000000	14.000000	5.700000
25%	106.000000	2.000000	2.000000	68.000000	66.000000	31.400000
50%	186.000000	2.000000	2.000000	143.000000	132.000000	42.900000
75%	312.000000	2.000000	3.000000	252.000000	218.000000	54.300000
max	835.000000	2.000000	4.000000	791.000000	656.000000	80.000000

Figure 1. Summary Statistics for all variables

## Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial phase in the analytical journey, especially when delving into the intricacies of healthcare-related datasets. This approach involves a systematic examination of the dataset to uncover patterns, relationships, and insights that lay the foundation for subsequent analyses. In the context of our healthcare data exploration, EDA serves as the initial step to gain a comprehensive understanding of the characteristics inherent in variables such as hospital stay duration ('Stay'), X-ray frequency ('Xray'), and infection risk ('InfctRsk').

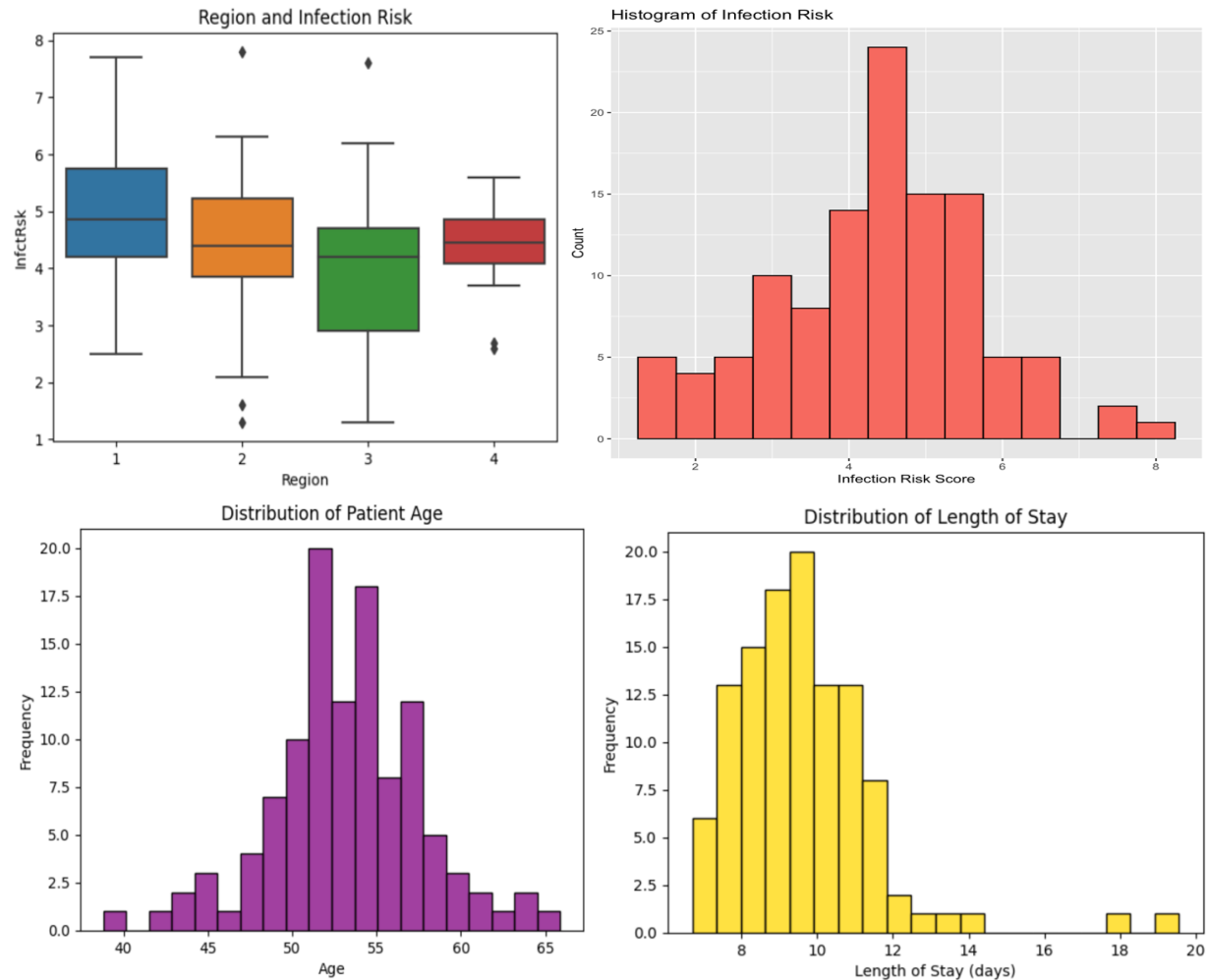


Figure 2. EDA Visualizations

To provide an understanding of the dataset we utilized the visualizations to illuminate the distribution and characteristics of the data by identifying patterns and potential outliers. The boxplot shows that Region 1 has the highest infection risk and all the regions have approximately the same median infection risk. Infection risk scores range from 1.3 to 7.8, with an average of about 4.35. The distribution of infection risk is fairly normal but shows a slight left skew. The number of X-rays administered ranges from 39.6 to 133.5, with an average around 81.63. The distribution is approximately normal. The length of hospital stays ranges from 6.7 to 19.56 days,

with an average of approximately 9.65 days. The distribution is slightly right skewed, indicating few longer stays.

## Data Analysis: Summary of Findings

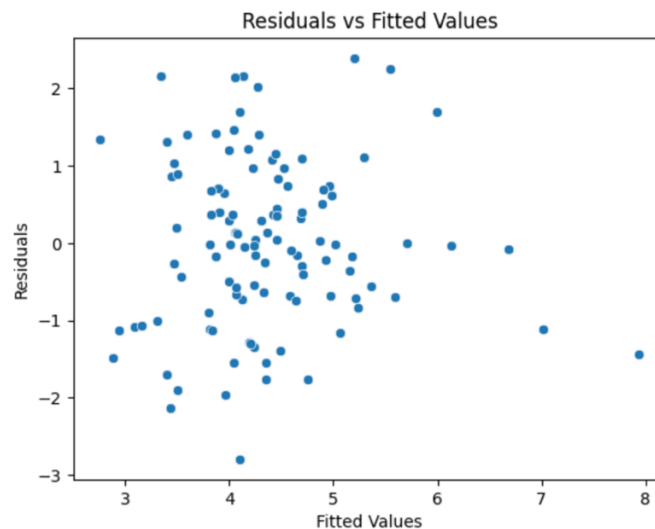
### Correlation Analysis

In an effort to unravel intricate patterns within our dataset, we employed correlation analysis and visualization techniques to discern relationships among variables. Utilizing a heatmap and pairplot, we systematically examined the correlations between key features and the target variable, 'InfctRsk.' Notably, our investigation revealed several noteworthy correlations: 'Stay' exhibited a significant positive correlation of 0.53 with infection risk, underscoring its role as a substantial predictor. Additionally, 'Xray' demonstrated a notable correlation of 0.45, while 'Age' displayed a marginal yet statistically significant correlation of 0.0011. Other features, such as 'Beds' (0.36), 'Census' (0.38), 'Nurses' (0.39), and 'Facilities' (0.41), also exhibited varying degrees of correlation with infection risk.

It is imperative to highlight that our later brute-force feature selection method identified 'Stay,' 'Age,' and 'Xray' as the optimal features, while the stepwise selection method emphasized 'Stay.' While correlation does not imply causation, these findings provide valuable insights into potential relationships that merit further exploration. The visual representation of these correlations through our heatmap and pairplot serves as a foundational step for comprehending the dataset's dynamics. Moving forward, additional analyses could delve deeper into the interplay of these features and their nuanced impacts on infection risk. It is crucial to acknowledge that correlation, though indicative of associations, does not establish a causal relationship. Therefore, caution must be exercised in interpreting these findings and exploring them within the broader context of the dataset and the specific domain under consideration. For details, please refer to the Appendix (Figure 2).

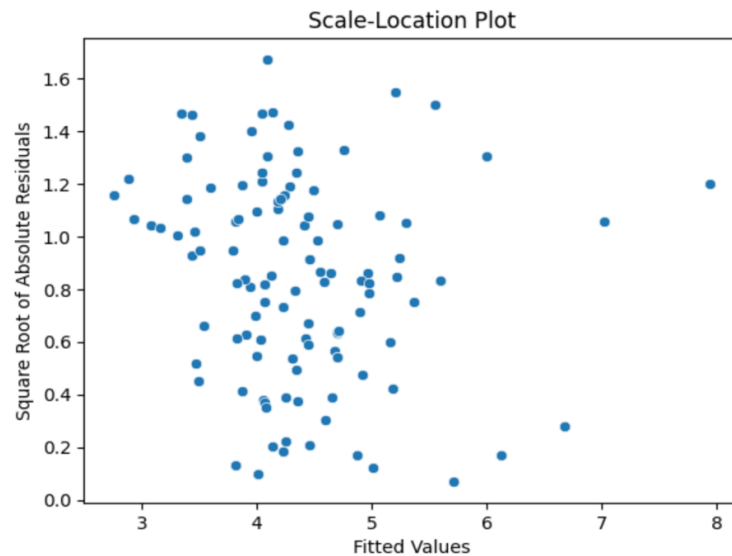
### Quality Check

#### 1.



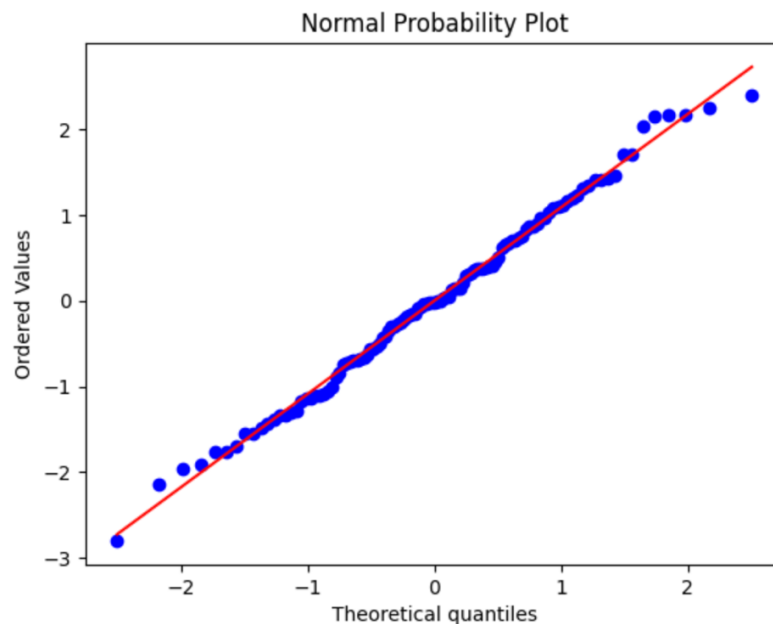
Ideally, residuals should be randomly distributed around zero as can be seen in the patterns and trends above.

## 2. Homoscedasticity



This plot helps assess whether the spread of residuals is consistent across different levels of the independent variable. The spread should ideally be constant as seen in the plot above.

## 3. Normality of Residuals



The plot mentioned in the context of assessing normality of residuals is typically a Normal Probability Plot, also known as a Q-Q (Quantile-Quantile) plot. In a Q-Q plot, the quantiles of the sample residuals are plotted against the quantiles of a theoretical normal distribution.

The points closely follow a straight line, it suggests that the residuals are approximately normally distributed.

#### 4. Multicollinearity

The Variance Inflation Factor (VIF) values indicate the presence of multicollinearity. Generally, a VIF greater than 10 is considered high, and a VIF around 5 or above may warrant further investigation.

Variable	VIF
const	32.88
Stay	1.17
Xray	1.17

In this case, the constant term ('const') has a high VIF, indicating potential multicollinearity.

#### Assumptions of Linear Regression

Linear regression is built upon several assumptions that, when met, ensure the reliability and validity of the model. It is crucial to be aware of these assumptions and evaluate their validity for the given dataset.

Linearity: The relationship between the predictors (e.g., 'Stay' and 'Xray') and the infection risk ('InfctRisk') is linear. For details, please refer to the Appendix (Figure 1).

1. Independence of Errors: The residuals (the differences between predicted and observed infection risk) are independent.
2. Homoscedasticity: The spread of the residuals is constant across all levels of the predictors.
3. Normality of Residuals: The residuals are normally distributed.
4. No Perfect Multicollinearity: The predictors, such as 'Stay' and 'Xray', are not perfectly correlated.

#### Hypothesis Testing

To understand the impact of how patients stay in the hospital (Stay) and how often they undergo X rays (Xray) on infection risk we conducted a hypothesis test to evaluate their contributions. Our null hypothesis (H0) stated that neither stay nor Xray variables were useful, in explaining the variability in infection risk while the alternative hypothesis (Ha) suggested that least one of these variables had power. The results of our analysis were quite compelling with an F statistic of 30.59 and a low p value of  $2.73 \times 10^{-11}$ . This provides evidence to support our findings. With a p value below the predetermined significance level ( $\alpha = 0.05$ ) we confidently reject the hypothesis. These results indicate that one variable—either Stay or Xray—plays a significant role in explaining the observed variations in infection risk. These findings emphasize the importance of both hospitals stay duration and X ray frequency when it comes to predicting infection risk thereby validating their significance within our framework.

During our investigation into the connection, between X ray frequency and infection risk while keeping hospital stay constant, we developed hypotheses to examine the significance of this relationship. The initial assumption (H0) suggested that the coefficient ( $\beta$ ) representing X ray frequency in our regression model was zero indicating no impact, on infection risk. In contrast the alternative assumption (Ha) stated that  $\beta XRay$  was not zero suggesting an influence

of X ray frequency on infection risk. Our statistical analysis yielded a p value of 0.000606. Considering a significance level ( $\alpha$ ) of 0.05 we observed that the p value was significantly lower leading us to reject the assumption. Therefore, we conclude that an increase in X ray frequency while keeping the number of days constant contributes to a rise in infection risk. This finding emphasizes the importance of considering X ray frequency as a determinant of infection risk within our analysis framework. The results offer insights for healthcare professionals and decision makers alike by highlighting the need for monitoring and strategic management of X ray procedures to minimize infection risks in healthcare settings.

During our investigation into the relationship, between age and infection risk we conducted a hypothesis test that carefully examined various aspects of regression analysis. The results we obtained indicate a linear relationship, with a slope of 0.0003285 and an intercept of 4.3374. The p value associated with the slope coefficient is 0.9908, which's important for hypothesis testing. Considering the hypothesis, which suggests no relationship (slope equals zero) and the alternative hypothesis, which suggests a relationship (slope not equal to zero) we decided to accept the null hypothesis due to the high p value of 0.9908. This means that there isn't evidence to support a linear relationship between 'Age' and 'InfctRsk' within our dataset's parameters. However, it's crucial to consider that statistical significance should be understood considering the study's objectives. This nuanced interpretation highlights the importance of aligning findings with the goals of our analysis while acknowledging any limitations, within our datasets scope. For details, please refer to the Appendix (Figure 3).

### **Model Fitting**

In our examination of how age may impact the likelihood of getting infected we conducted a linear regression analysis to understand the nature of this relationship within our dataset. The results of this analysis showed that there is a connection, between age and infection risk as evidenced by the negligible slope associated with age. Additionally, we found that the p value, a measure was not significant. This means that any observed link between age and infection risk is more likely due to variation than a statistically meaningful pattern. Therefore, based on our analysis age does not appear to be a predictor of infection risk in this dataset. The high p value suggests that changes in age are not consistently accompanied by corresponding changes in infection risk. While this finding doesn't completely rule out the influence of age it highlights that relying on age as a predictor may have limited effectiveness in explaining variations in infection risk. It's important to interpret these findings considering the context of our dataset and acknowledge that other factors may play roles in predicting infection risk within this specific domain.

We utilized a regression model called M1. Incorporated key factors such, as 'Stay' and 'Xray' to predict the level of infection risk ('InfctRsk'). The performance of our model was thoroughly evaluated using the error (MSE) which is a reliable metric commonly used for assessing regression models. Our analysis revealed that the calculated MSE was 1.189 indicating the squared difference, between the predicted values of our model and the actual infection risk values. This result suggests a level of accuracy in our model as lower MSE values generally indicate more precise predictions. By focusing on factors like 'Stay' and 'Xray' our linear regression model effectively captured patterns in the dataset that are related to infection risk. Although the MSE provides a measure of performance, exploration and potential comparison



with alternative models could offer deeper insights into how well our approach predicts infections.

To gain an understanding of the factors influencing infection risk ('InfctRsk') we expanded our analysis beyond the initial features 'Stay' and 'Xray' by incorporating additional variables such as 'Age' 'Beds,' 'Census,' 'Nurses' and 'Facilities,' into a multiple linear regression framework. This extension aimed to capture a view of these influencing factors. Through both selection and a brute force approach our analysis shed light on the relevance of these additional features. Interestingly when we used the selection method, we found that 'Stay' was the most influential predictor, in predicting infection risk. This was further supported by the brute force selection method, which consistently identified 'Stay' and 'Xray' as the features. Surprisingly including variables like 'Age' 'Beds,' 'Census,' 'Nurses' and 'Facilities' didn't significantly improve our model's performance. Instead, it became clear that 'Stay' played a role in capturing patterns within the dataset. This emphasizes how important both 'Stay' and 'Xray' are in predicting infection risk with factors considered.

Additionally, we introduced an interaction term ('Stay' \* 'Xray') to explore any synergies or dependencies between these two features in predicting infection risk (denoted as "InfctRsk") in Model M2. The subsequent analysis of this updated model provided insights, into the impact of this interaction term. However, the results showed that the interaction term didn't have an effect, on the risk of infection. This is evident from the coefficients and p values associated with it. The regression analysis yielded an intercept of 3.0924 a coefficient of 0.607 for 'Stay' 0.052 for 'Xray '. A cross term coefficient of 0.0033 for the interaction between 'Stay' and 'Xray.' While the coefficients for 'Stay' and 'Xray' were statistically significant the p value of 0.195 for the interaction term indicated that it didn't contribute significantly to the power of the model. This conclusion is further supported by observing that introducing the interaction term didn't lead to an improvement in capability as there was no notable change, in the R squared value.

To enhance Model M2's ability we performed an analysis to evaluate how excluding the 'Age' variable impacted its estimation of infection risk. We modified Model M2 accordingly by removing this variable and then assessed its resulting R value. Surprisingly we found that this modified model had an R value around 0.436 indicating a slight difference compared to the original model. As a result, our findings suggest that excluding 'Age' did not have an impact, on the model's ability to explain the variation in infection risk. This discovery highlights the strength of Model M2, which is primarily influenced by 'Stay' and 'Xray' in capturing the dynamics that affect infection risk.

It is important to consider the implications of our analysis and emphasize the tradeoffs between model simplicity and interpretability. Although leaving out 'Age' did not greatly affect how well the model performed it does prompt us to reflect on finding a balance between having a model and making it easier to understand. This perspective is especially relevant in healthcare settings where decision makers value models that're also easy to comprehend.

When comparing two regression models, M1 and M2 we evaluated their performance using measures. M1, which includes predictors like 'Stay' and 'Xray' showed an R value of 0.357 indicating its ability to explain around 35.7% of the variability, in the variable. On the hand M2,

incorporating an interaction term ('Stay:Xray') demonstrated a slightly higher R squared value of 0.367. Despite this improvement, the Adjusted R squared, which considers model complexity showed a decrease, in power for M2 (0.346 for M1 and 0.350 for M2). The F statistic, which assesses the significance of the model favored M1 with a value of 30.59 compared to 21.09 for M2. Therefore, our analysis suggests that the interaction effect between 'Stay' and 'Xray' may not have a role in influencing infection risk highlighting the importance of 'Stay' and 'Xray' as influential predictors, in our linear regression model.

For all the models, please refer to the Appendix (Figure 4-8).

### **Implementation**

The implementation of Model M1 involves using a regression approach, where we consider the factors 'Stay' and 'Xray' to predict the infection risk labeled as 'InfctRsk'.

To predict the infection risk, for a patient we can utilize the equation:

$$InfctRsk = 0.1506 + 0.2958 \times Stay + 0.0202 \times Xray.$$

For instance, let's say we want to estimate the infection risk for a patient who stays in the hospital for 10 days and has undergone 5 X-rays. By substituting these values into the equation mentioned above we find that their projected infection risk is 2.91. Therefore, according to our regression model, a patient, with a stay of 10 days and having had 5 X rays is expected to have an infection risk of 2.91.

For details, please refer to the Appendix (Figure 9).

## Recommendations for Healthcare Providers

As we wrap up our comprehensive analysis of the healthcare dataset we have discovered important findings and insights that may offer valuable guidance, for healthcare providers.

1. **Emphasize the Duration of Hospital Stay:**  
We consistently observed that the length of a patients stay in the hospital ('Stay') is an indicator of infection risk. Healthcare providers are encouraged to take note of this correlation and consider implementing protocols to monitor and manage stays. Doing so can contribute to infection prevention strategies.
2. **Utilize Insights from X ray Frequency:**  
Our analysis has revealed a connection between the frequency of X ray examinations ('Xray'). Infection risk. Healthcare providers should consider incorporating this information into risk assessments as it can help in detection and preventive measures against infections.
3. **Improving Healthcare Facilities:**  
Our analysis has uncovered an inverse relationship between the quality of healthcare facilities and infection risks. We encourage healthcare providers to prioritize investments in enhancing hospital infrastructure aiming for improvements in safety. Initiatives such as facility upgrades, modernization efforts and strict adherence to sanitation standards can collectively create an environment, for healthcare delivery.
4. **Considering Regional Factors:**  
The discovery of higher infection risks, in hospitals on the West Coast highlights the importance of taking region factors into account when it comes to infection control. It is essential for healthcare providers in this region to thoroughly review their infection control protocols and practices. Adapting strategies to address risk factors whether influenced by demographics, climate or other variables will play a role in reducing infection risks and creating a safer healthcare environment.
5. **Easy to understand Overview for Patients:**  
Our data driven analysis offers insights for both healthcare providers and patients alike. For healthcare facilities this report serves as a resource for optimizing patient care particularly focusing on facility related issues and regional risk factors. As for patients this report equips them with knowledge that can guide their healthcare journeys in a manner. We encourage patients to engage with their healthcare providers inquire about improvements at the facilities they visit and stay informed about risk factors. This patient friendly overview aims to empower individuals so they can make informed decisions regarding their healthcare experiences.

By implementing these recommendations and continuously monitoring and adapting practices we can contribute to an improvement, in safety and the overall quality of healthcare delivery.

## Conclusion and Future Directions

In conclusion, our comprehensive analysis of the healthcare dataset has provided valuable insights into the factors influencing infection risk within healthcare settings. The initial questions posed at the introduction were systematically addressed, shedding light on the significance of variables such as hospital stay duration ('Stay') and X-ray frequency ('Xray') in predicting infection risk ('InfctRsk'). The linear regression model underscored the relevance of these variables, with 'Stay' consistently emerging as a key predictor. Subsequent feature selection methods corroborated the importance of 'Stay,' further supported by correlation analyses. Moreover, the comparison of two models highlighted the nuanced trade-off between model simplicity and explanatory power. Recommendations for healthcare providers emphasize the importance of monitoring the hospital stay duration and X-ray frequency, leveraging predictive models, and continuous adjustment of protocols. Looking forward, opportunities for future work include exploring the interplay of additional factors and refining predictive models for enhanced infection risk management. This analysis sets the stage for a more targeted and data-driven approach to infection prevention within healthcare settings, emphasizing the continual evolution and adaptation of strategies to ensure patient safety. As patient safety remains a dynamic and evolving challenge, these directions will contribute to a more targeted and adaptive approach in infection risk management.

## Appendix

### 1. Linearity Assumption

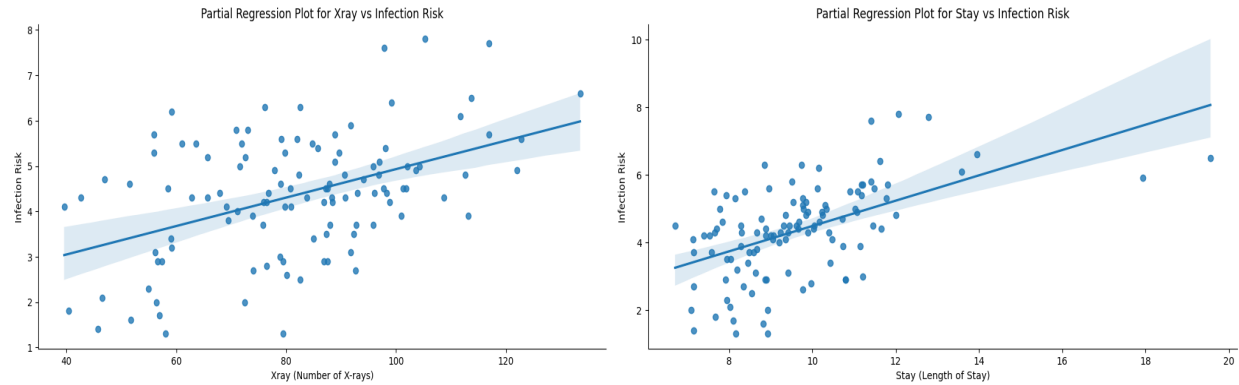


Figure 1: Linearity Assumption

### 2. Correlation Heatmap and Visualizations:

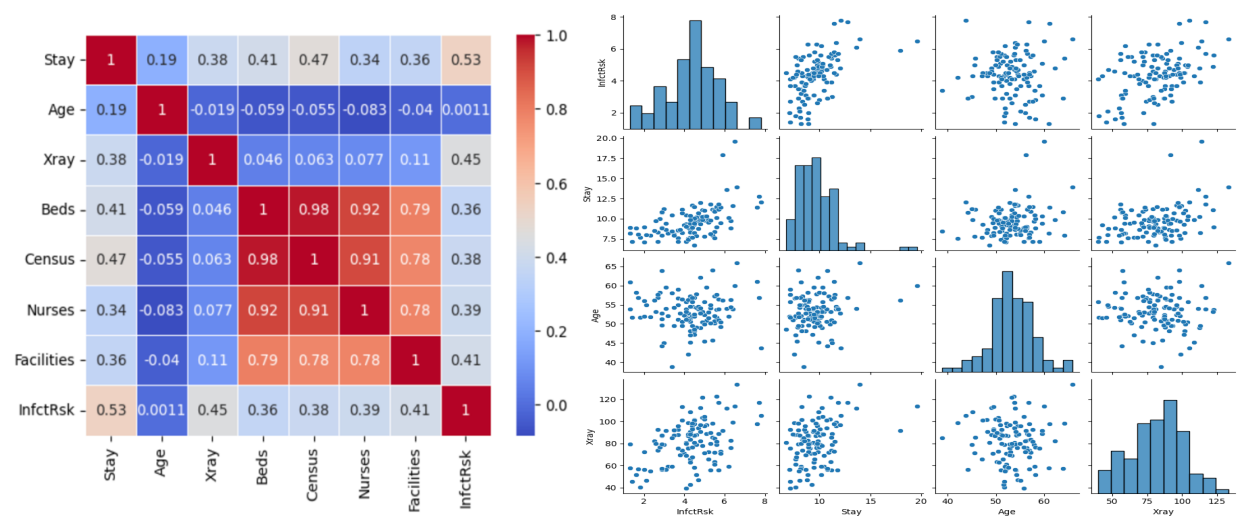


Figure 2: Correlation Analysis

### 3. Age and Infection Relationship

```
1 from scipy.stats import linregress
2
3 slope, intercept, r_value, p_value, std_err = linregress(df['Age'], df['InfctRsk'])
4 print(f'Slope: {slope}, Intercept: {intercept}, p-value: {p_value}')
```



Slope: 0.00032854419080125344, Intercept: 4.337378238791967, p-value: 0.990831460298162

Figure 3: Age and Infection Relationship

#### 4. Final Model M1:

OLS Regression Results						
=====						
Dep. Variable:	InfctRsk	R-squared:	0.357			
Model:	OLS	Adj. R-squared:	0.346			
Method:	Least Squares	F-statistic:	30.59			
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	2.73e-11			
Time:	06:49:18	Log-Likelihood:	-168.00			
No. Observations:	113	AIC:	342.0			
Df Residuals:	110	BIC:	350.2			
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	-0.1506	0.585	-0.257	0.797	-1.310	1.009
Stay	0.2958	0.058	5.098	0.000	0.181	0.411
Xray	0.0202	0.006	3.531	0.001	0.009	0.032
=====						
Omnibus:	0.653	Durbin-Watson:	1.874			
Prob(Omnibus):	0.721	Jarque-Bera (JB):	0.744			
Skew:	0.032	Prob(JB):	0.690			
Kurtosis:	2.608	Cond. No.	485.			
=====						

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

*Figure 4: Final Model M1*

#### 5. Feature Selection Results:

Method:	Least Squares		F-statistic:		21.09	
Date:	Mon, 11 Dec 2023		Prob (F-statistic):		7.56e-11	
Time:	06:49:15		Log-Likelihood:		-167.12	
No. Observations:	113		AIC:		342.2	
Df Residuals:	109		BIC:		353.2	
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-3.0924	2.329	-1.328	0.187	-7.708	1.523
Stay	0.6070	0.245	2.474	0.015	0.121	1.093
Xray	0.0520	0.025	2.080	0.040	0.002	0.101
Stay:Xray	-0.0033	0.003	-1.305	0.195	-0.008	0.002
Omnibus:	0.331	Durbin-Watson:		1.846		
Prob(Omnibus):	0.847	Jarque-Bera (JB):		0.497		
Skew:	0.077	Prob(JB):		0.780		
Kurtosis:	2.714	Cond. No.		1.98e+04		

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 1.98e+04. This might indicate that there are strong multicollinearity or other numerical problems.

*Figure 5: Model M2*

```

1 from itertools import combinations
2 from sklearn.metrics import mean_squared_error
3 from sklearn.model_selection import train_test_split
4
5 # Select predictor variables
6 X = df[['Stay', 'Age', 'Xray', 'Beds', 'Census', 'Nurses', 'Facilities']]
7
8 # Response variable
9 y = df['InfctRsk']
10
11 # Split the data into training and testing sets
12 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
13
14 best_features = None
15 best_mse = float('inf')
16
17 # Try all possible combinations of features
18 for k in range(1, len(X.columns) + 1):
19     for combo in combinations(X.columns, k):
20         model = LinearRegression()
21         model.fit(X_train[list(combo)], y_train)
22         y_pred = model.predict(X_test[list(combo)])
23         mse = mean_squared_error(y_test, y_pred)
24
25         if mse < best_mse:
26             best_mse = mse
27             best_features = combo
28 best_features
29 #multi colinearity

```

('Stay', 'Age', 'Xray')

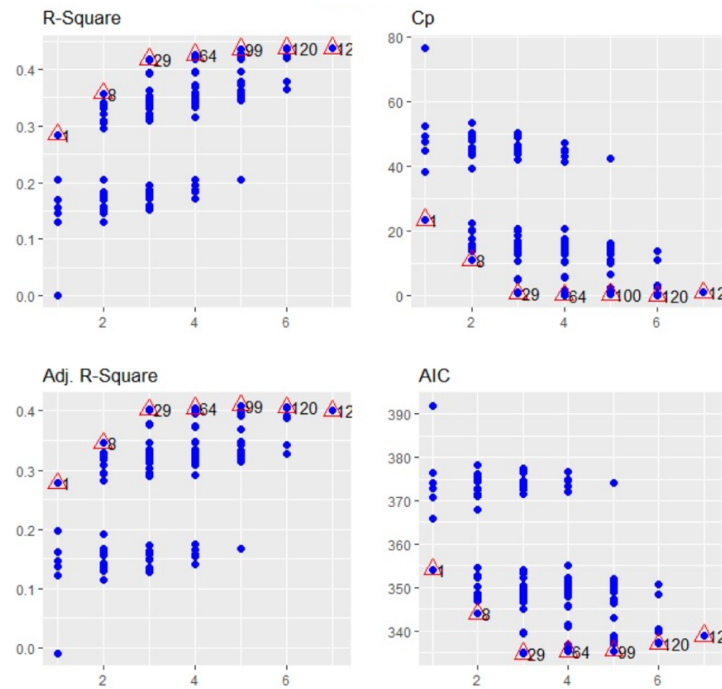


Figure 6: Brute-force Selection Output

```

1 from sklearn.feature_selection import SequentialFeatureSelector
2 from sklearn.linear_model import LinearRegression
3 from sklearn.model_selection import cross_val_score
4
5 # Select predictor variables
6 X = df[['Stay', 'Age', 'Xray', 'Beds', 'Census', 'Nurses', 'Facilities']]
7
8 # Response variable
9 y = df['InfctRisk']
10
11 # Forward stepwise feature selection
12 model = LinearRegression()
13 sfs = SequentialFeatureSelector(model, n_features_to_select=1, direction='forward', scoring='neg_mean_squared_error',
14 sfs.fit(X, y)
15
16 # Display selected features
17 selected_features = list(X.columns[list(sfs.get_support())])
18 selected_features

```

['Stay']

*Figure 7: Forward Stepwise Selection Output*

```

1 # Fit Model M2 without 'Age'
2 X_m2_without_age = df[['Stay', 'Xray', 'Beds', 'Census', 'Nurses', 'Facilities']]
3 model_m2_without_age = LinearRegression()
4 model_m2_without_age.fit(X_f2_without_age, y)
5
6 # Make predictions
7 y_pred_m2_without_age = model_m2_without_age.predict(X_m2_without_age)
8
9 # Calculate R-squared
10 r2_f2_without_age = r2_score(y, y_pred_f2_without_age)
11
12 print(f"R-squared for Model F2 without 'Age': {r2_f2_without_age}")

```



R-squared for Model F2 without 'Age': 0.435975495659802

*Figure 8: Model M2 without 'Age'*

## 6. Implementation

```

1 import numpy as np
2
3 # Coefficients
4 intercept = -0.1506
5 coef_stay = 0.2958
6 coef_xray = 0.0202
7
8 # Sample input data
9 stay_value = 10 # Replace with the actual length of stay
10 xray_value = 5 # Replace with the actual number of X-rays
11
12 # Calculate predicted InfctRisk
13 predicted_infct_risk = intercept + coef_stay * stay_value + coef_xray * xray_value
14
15 print(f'Predicted Infection Risk: {predicted_infct_risk}')
16

```

Predicted Infection Risk: 2.9084000000000003

*Figure 9: Implementation*