Answer to Question 1:

In addressing the missing values in the patient age data, I would consider several techniques based on the nature of the data and the amount missing. Since 15% of the data is missing, which is significant but not overwhelming, and given the size of the dataset, I would employ a method that could provide a good balance between bias and variance while maintaining the integrity of the dataset.

Step-by-Step Approach:

1. Data Analysis:
   - First, I would conduct an exploratory data analysis to understand the distribution of the age data. This would include identifying patterns or relationships between age and other variables.

2. Choosing the Imputation Technique:
   - Option A: Multiple Imputation - Given that this is healthcare data, using multiple imputation can be effective as it handles missing data by imputing multiple values to create several complete datasets. It then averages the results to account for the uncertainty of the missing data.
   - Option B: Model-Based Imputation - If there are strong correlations between age and other variables, using regression techniques to predict missing ages can be a viable option. This approach would leverage the relationships within the data to provide a statistically grounded imputation.

3. Validation:
   - After imputation, it's crucial to validate the results by checking the consistency of the imputed ages with known age distributions in similar datasets or populations. This could involve statistical tests or visual inspections of the data distribution.

4. Handling Missing Diagnosis Information:
   - If critical data such as diagnosis information were missing, a more cautious approach would be necessary. Given its importance, consultation with domain experts and possibly acquiring additional data might be preferable to statistical imputation. If imputation is still considered, sophisticated machine learning models that can incorporate the uncertainty and complexity of healthcare data, such as Random Forest or Neural Networks, might be appropriate.

Justification:
The choice of multiple imputation and model-based techniques is guided by the desire to preserve the statistical relationships in the data without reducing the sample size. This approach is beneficial in a large dataset where the integrity of relationships is crucial for subsequent analyses.

Rubric for Evaluating the Answer:

1. Understanding of Data Imputation Techniques (30 Points):
   - Demonstrates a clear understanding of different data imputation methods.
   - Explains why specific techniques are preferred based on the dataset characteristics.

2. Application of Statistical Knowledge (30 Points):
   - Appropriately applies statistical reasoning to justify the choice of imputation method.
   - Discusses how the chosen method handles the bias-variance trade-off.

3. Problem Solving and Innovation (20 Points):
   - Provides a creative yet practical solution to handle missing values in a critical healthcare dataset.
   - Considers the use of advanced techniques for more critical missing data, reflecting an understanding of the domain-specific implications.

4. Validation and Ethical Considerations (20 Points):
   - Highlights the importance of validating the imputed data to ensure accuracy and reliability.
   - Addresses potential ethical considerations, especially when dealing with sensitive health data, ensuring the solution does not mislead or result in poor clinical decisions.

Total Score: 100 Points

Answer to Question 2:

For the task of handling outliers in a financial dataset used for forecasting stock prices, especially when these outliers correspond to significant market news days, a nuanced approach is needed. The primary goal is to preserve the integrity of the data while ensuring the predictive model reflects true market behaviors without being skewed by exceptional cases.

Step-by-Step Approach:

1. Outlier Detection:
   - Z-Score Analysis: This method helps identify how many standard deviations an observation is from the mean. In financial data, observations beyond 3 standard deviations are typically considered outliers. This method is sensitive to the mean and standard deviation, which can be heavily influenced by extreme values themselves.
   - Interquartile Range (IQR): This method involves calculating the first (Q1) and third quartiles (Q3) and identifying data points outside the range [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]. This method is less sensitive to extreme values and thus more robust for financial datasets.

2. Deciding on Treatment:

- Adjustment: For data points identified as outliers that coincide with significant market news, instead of removing these points, adjusting them may be more appropriate. For instance, applying a smoothing technique like moving averages or a moderate dampening factor can integrate the influence of the event without letting it dominate the dataset.
- Preservation for Model Training: In cases where the outliers are legitimate (reflecting true market reactions to significant events), they should be preserved within the training dataset. This inclusion ensures that the model can learn the impact of similar events in the future.

3. Impact Analysis:
- Assess the effect of outliers on model performance by comparing metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) before and after treating the outliers.
- Conduct backtesting with and without outlier adjustments to observe the impact on the model's predictive accuracy and overfitting tendencies.

Justification:
The choice to adjust rather than remove outliers related to significant news events respects the nature of financial markets, where such events can drastically affect stock prices. Smoothing or adjusting these outliers ensures that their impact is recognized but not overstated, aiding in developing a more robust and realistic forecasting model.

Rubric for Evaluating the Answer:

1. Analytical Rigor (30 Points):
- Clearly identifies appropriate techniques for outlier detection, explaining the merits and limitations of each (e.g., Z-score and IQR).
- Demonstrates understanding of financial data characteristics and the effects of outliers.

2. Approach to Outlier Management (30 Points):
- Offers a well-reasoned strategy for dealing with outliers, distinguishing between errors and legitimate extreme values.
- Discusses methods for adjusting or preserving outliers in the context of their impact on predictive modeling.

3. Modeling Impact Assessment (20 Points):
- Describes methods for evaluating the impact of outlier treatment on model performance, including statistical tests and backtesting.
- Provides criteria for deciding when and how to incorporate outliers into model training.

4. Critical Thinking and Innovation (20 Points):
- Applies innovative thinking to the problem of managing outliers in stock price data, suggesting methods like smoothing adjustments that reflect real-world trading scenarios.
- Shows foresight in how these decisions might affect model reliability and business decisions.

Total Score: 100 Points

Answer to Question 3:

When preparing a dataset for a real estate price prediction model that includes features like property size in square feet and local crime rate per 1,000 residents, choosing between normalization and standardization depends on the nature of the data and the requirements of the machine learning algorithms to be used.

Step-by-Step Approach:

1. Analysis of Data Characteristics:
   - First, assess the distribution of each feature. If the property size and crime rate data follow a normal distribution, standardization may be appropriate. If the scales of the features vary widely and do not follow a Gaussian distribution, normalization might be more suitable.

2. Choosing the Technique:
   - Normalization: This technique rescales the data to a fixed range, typically 0 to 1. This is particularly useful when dealing with features that do not follow a normal distribution. Since algorithms like k-nearest neighbors (KNN) and neural networks are sensitive to the scale of the input data, normalization helps ensure that all features contribute equally to the distance calculations, preventing features with larger scales from dominating.
   - Standardization: This method rescales data to have a mean of 0 and a standard deviation of 1, forming a Gaussian distribution. It is suitable for models like linear regression, logistic regression, and support vector machines that assume data is normally distributed.

3. Application of Techniques:
   - Apply normalization to the property size, as the range can vary significantly between properties, potentially skewing the model if not adjusted. Normalize the crime rate if the data shows extreme values or if the algorithm requires scaled data.
   - Standardize both features if the chosen model assumes a normal distribution of input data, such as in linear models.

4. Evaluation:
   - After applying the chosen technique, evaluate the model's performance. Check for improvements in prediction accuracy and generalization capability. Look for signs of overfitting or underfitting, which can indicate whether the data transformation was beneficial or if adjustments are needed.

Justification:
Normalization is chosen for its ability to scale data uniformly, ensuring that no single feature overwhelms others in distance-based algorithms. Standardization is used when models require

normally distributed data, helping to maintain consistency in input data and improving algorithm efficiency.

Rubric for Evaluating the Answer:

1. Understanding of Data Preprocessing Techniques (30 Points):
   - Demonstrates comprehensive knowledge of both normalization and standardization, including their mathematical foundations and effects on data.
   - Clearly differentiates when each technique is preferable based on the statistical properties of the data and the algorithms used.

2. Practical Application (30 Points):
   - Provides a detailed explanation of how to apply normalization and standardization to the real estate dataset.
   - Discusses the reasons for choosing one technique over the other for specific features within the dataset.

3. Model Performance Evaluation (20 Points):
   - Describes how to assess the impact of data preprocessing on the machine learning model's performance.
   - Suggests metrics or methods to evaluate the effectiveness of the preprocessing in improving model accuracy and handling overfitting or underfitting.

4. Critical Analysis and Decision Making (20 Points):
   - Critically analyzes the potential impacts of each preprocessing technique on the model's learning process and prediction accuracy.
   - Makes informed decisions based on the characteristics of the dataset and the requirements of the predictive model, showing an ability to adapt techniques to fit specific needs.

Total Score: 100 Points