Report on Synthetic Nurse Pay Data Analysis

Table of contents

1. Introduction

1.1 Objectives

* 1. Scope

1. Data Overview

2.1 Data

2.2 Data Snapshot

1. Data Preprocessing

**3.1 Handling Missing Values**

**3.2 Converting dates**

3.3 Encoding Categorical Variables

1. Exploratory Data Analysis

**4.1 Distribution of Hourly Pay Rate**

4.2 Contract Length Analysis

4.3 Correlation Analysis

1. Model Development

5.1 Model Selection

5.2 Training and Evaluation

1. Model Performance

6.1 Discussion

6.2 Analysis

1. Findings and Insights
2. Conclusion
3. Future Work
4. **Introduction**

In recent years, the healthcare industry has faced numerous challenges, particularly in staffing and compensation for nurses. Accurate data on nurse pay rates is critical for hospital administration and policy makers to ensure competitive compensation, retain skilled nurses, and provide quality patient care. This report aims to analyze synthetic nurse pay data generated with the help of ChatGPT. We will explore the factors influencing nurse pay rates, develop predictive models to estimate pay rates, and provide actionable insights based on our findings.

This data contain hourly pay rates for nurses across the major metros in the US as specified. It ensures seasonal upticks during flu and holiday seasons, randomizes contract start and end dates within the given range, and includes necessary details such as job title, location, hospital name, and pay rate.

1.1 Objectives

* To understand the distribution and factors affecting nurse pay rates.
* To develop predictive models for estimating nurse pay rates based on various features.
* To provide recommendations based on the analysis.

1.2 Scope

This report includes detailed data preprocessing steps, exploratory data analysis (EDA), model development, and evaluation. It also discusses alternative approaches and methodologies that could be employed in such analyses.

**2. Data Overview**

2.1 Data

The dataset used for this analysis is synthetic nurse pay data generated with the assistance of ChatGPT. The dataset comprises the following features:

* **Job Title**: The specific role or title held by the nurse (e.g., Registered Nurse, Nurse Practitioner).
* **City**: The city where the nurse is employed.
* **State**: The state where the nurse is employed.
* **Hospital Name**: The name of the hospital employing the nurse.
* **Contract Start Date**: The start date of the nurse's contract.
* **Contract End Date**: The end date of the nurse's contract.
* **Hourly Pay Rate**: The hourly compensation rate for the nurse.

2.2 Data Snapshot

A preview of the dataset is shown below:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Job Title** | **City** | **State** | **Hospital**  **Name** | **Contract**  **Start**  **Date** | **Contract**  **end**  **Date** | **Pay rate** |
| 1 | Registered Nurse Surgery | Houston | TX | Houston Govt | 2023-11-02  15:36:11 | 2023-11-02  15:36:11 | 64.27 |
| 2 | Registered Nurse Pediatric | San Diego | CA | San Diego  Community | 2023-04-30 21:14:41 | 2023-04-30 21:14:41 | 30.39 |
| 3 | Registered Nurse Telemetry | San Diego | CA | San Diego  Community | 2024-10-10 10:42:34 | 2024-10-10 10:42:34 | 45.63 |
| 4 | Lab Technician | San Diego | CA | San Diego  NonProfit | 2023-07-07 07:10:11 | 2023-07-07 07:10:11 | 48.6 |
| 5 | Registered Nurse MedSurg | Dallas | TX | Houston Govt | 2024-09-06 08:58:53 | 2024-09-06 08:58:53 | 40.54 |

The dataset contains 2,50,000 records, providing a rich source of information for analysis.

### 3. Data Preprocessing

Data preprocessing is a critical step in ensuring the dataset is clean and suitable for analysis. It involves handling missing values, converting data types, feature engineering, encoding categorical variables, and scaling numerical features.

**3.1 Handling Missing Values**

First, we check for any missing values in the dataset. Missing data can lead to biases and affect model performance. In this dataset, there are no missing values, ensuring data integrity.

**3.2 Converting Dates**

The Contract Start Date and Contract End Date columns are converted to datetime format to facilitate calculations involving dates. This conversion allows us to create a new feature, Contract month, which represents the duration of the nurse's contract in days.

3.3 Encoding Categorical Variables

Categorical variables such as **Job Title, City, State**, and **Hospital Name** are encoded using binary encoding. This technique converts categorical variables into a format suitable for machine learning algorithms. Ordinal data such as **Specialization, Season,** and **Desirability** are encoded using Label encoding.

**4. Exploratory Data Analysis**

EDA helps us understand the underlying patterns and relationships in the data. It involves visualizing distributions, identifying correlations, and examining group differences.

**4.1 Distribution of Hourly Pay Rate**

The distribution of hourly pay rates is analyzed to understand the central tendency and spread of nurse compensation.

4.2 Contract Analysis

We analyze the distribution of contract lengths to understand typical contract durations for nurses.

4.3 Correlation Analysis

Correlation analysis helps identify relationships between numerical features. We calculate the Pearson correlation for linear positive relationship between attributes and kendel coorelation for negative relationship between features.

**5. Model Development**

We develop predictive models to estimate nurse pay rates. Several algorithms are considered, each with its strengths and weaknesses.

5.1 Model Selection

* Linear Regression: A simple and interpretable model. it is effective predicting continuous target variables and provide coefficient that explains the linear relationship between features and the target variables.
* Random Forest Regressor: An ensemble of decision trees that improves robustness and accuracy by averaging multiple trees. To address the limitations of non linearity relationship in features.

5.2 Training and Evaluation

The dataset is split into training and testing sets to evaluate model performance. We use metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (MSE)

* MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between predicted and actual observations.
* RMSE measures the square root of the average of the squared differences between predicted and actual values. It gives a relatively high weight to large errors, meaning that RMSE is more sensitive to outliers than MAE.

### 6. Model Performance

We compare the performance of the different models based on the evaluation metrics.

6.1 Discussion

* Linear Regression: Provides a baseline with reasonable performance but may not capture complex relationships.
* Random Forest: Offers the best performance by averaging multiple trees, reducing overfitting, and improving generalization.

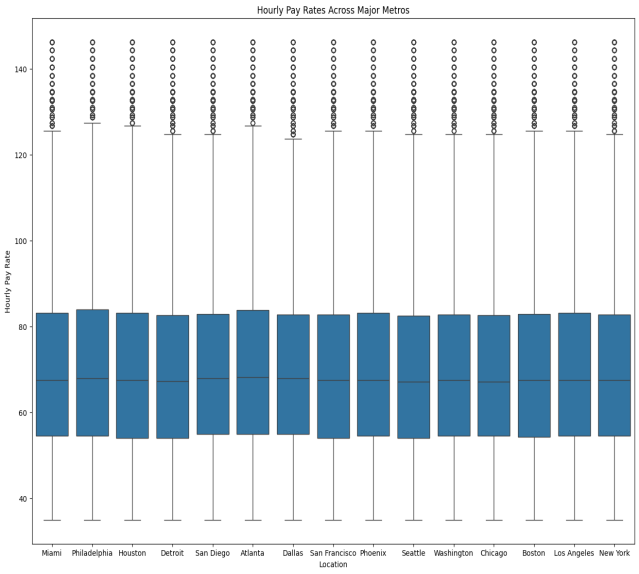
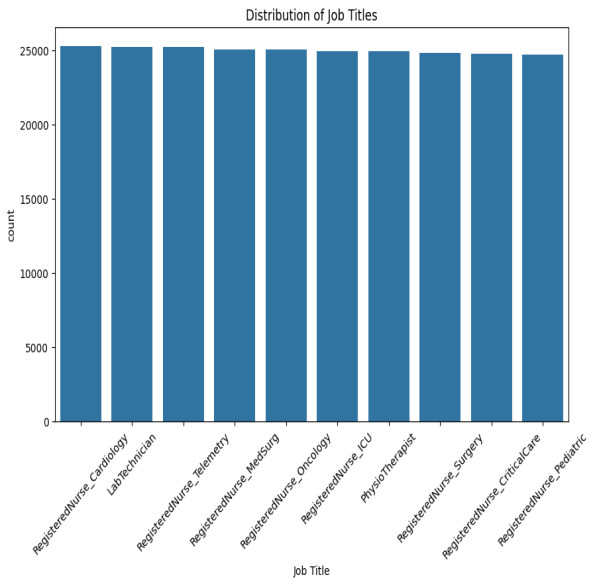
6.2 Analysis

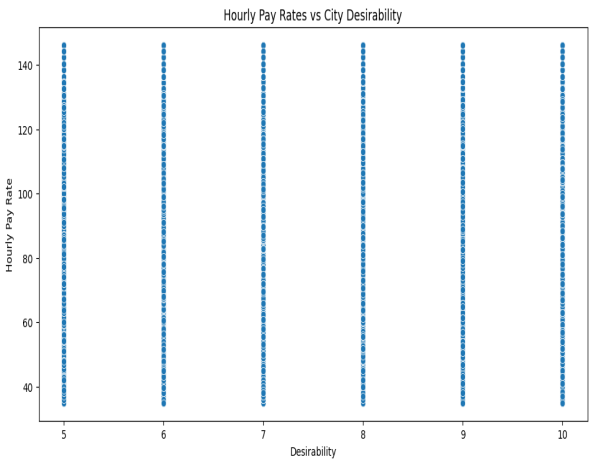
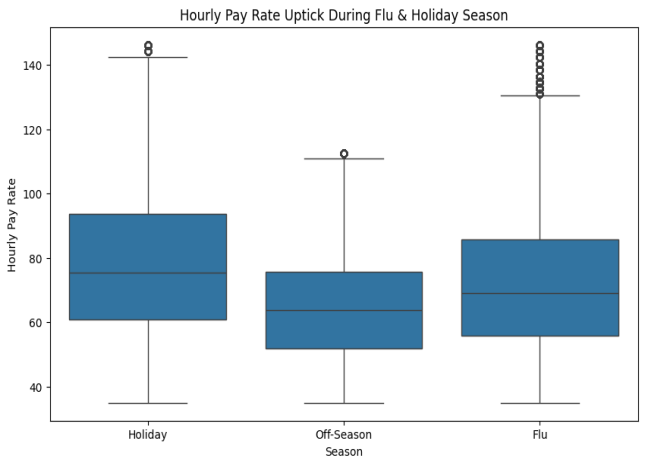
* PCA : PCA is used for dimensionality reduction while retaining most of the variance in the data.
* Variance : variance refers to how much the model's predictions change when different training data is used.
* **Feature Importance:** Random Forest models can also provide insights into feature importance, indicating which features have the most significant impact on the predictions.

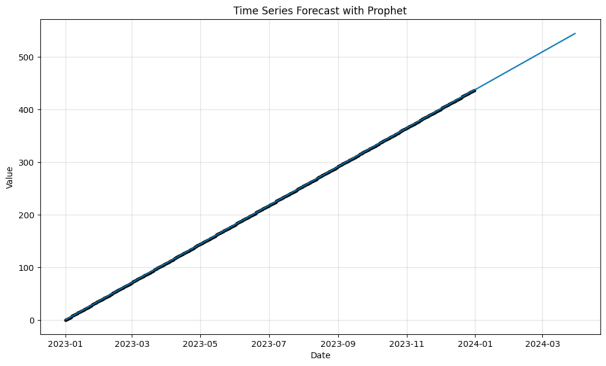
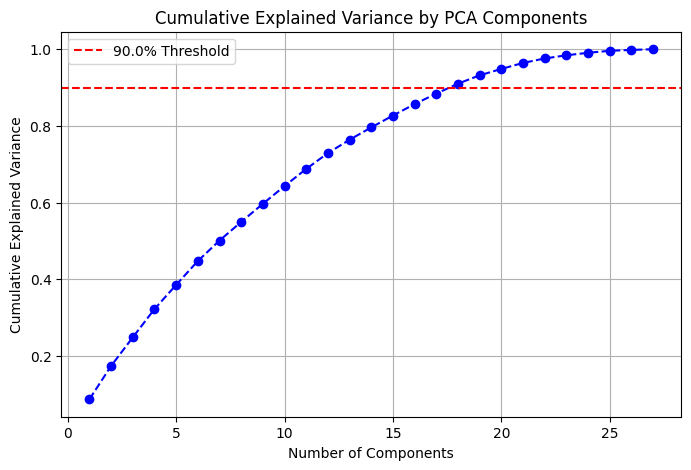
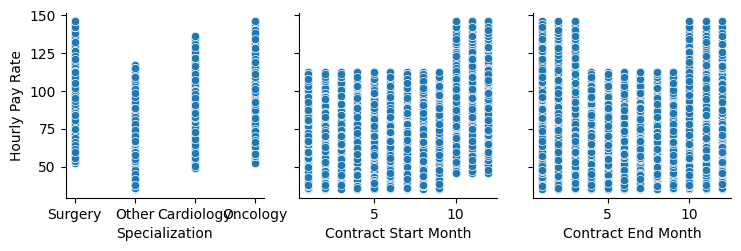
**7. Findings and Insights**

The analysis of synthetic nurse pay data reveals several key insights:

* Count of nurses working on different job titles : Plotted Bar graph that shows data is equally distributed under job titles.
* Hourly pay rates across major metros : Plotted boxplot showing hourly pay rates across major metropolitan areas most cities, the IQR falls between approximately 60 and 80 dollars per hour, indicating where the bulk of the pay rates lie
* Hourly payrates during flu and holiday season : the hourly pay rates during flu and holiday season is higher than the other months.
* Job Titles: Different nursing roles have distinct pay scales, with specialized positions like Nurse Practitioners earning higher rates.
* Prophet : **Prophet** is a forecasting tool designed for making time-series predictions, especially for data with strong seasonal effects and multiple seasons.







**8. Conclusion**

This report provides a comprehensive analysis of synthetic nurse pay data. By preprocessing the data, conducting EDA, and developing predictive models, we gain valuable insights into the factors influencing nurse pay rates. The Random Forest model offers the best performance, accurately estimating hourly pay rates based on various features.

**9. Future Work**

Future analyses could explore additional factors influencing nurse pay rates, such as:

To enhance model performance in the future, focus on advanced feature engineering and selection, rigorous hyperparameter tuning using grid or random search, and exploring ensemble methods like stacking and boosting.

By expanding the scope of the analysis, we can gain deeper insights and make more informed recommendations.

THANKYOU