**Final Project Report**

**Group 1**

Project Name: Compassionate Decision Support for Patients & Families to Choose Care Services

Anusha Basam

Aoi Ikemori

Bhavana Chittineni

Vandana Kolla

**University of North Texas**

DSCI 5260: Business Process Analytics

Maryam Khatami, Ph.D.

April 27, 2025

**Table of Contents**

[List of Figures 3](#_Toc196682529)

[List of Tables 4](#_Toc196682530)

[1. Abstract 5](#_Toc196682531)

[2. Introduction 5](#_Toc196682532)

[3. Literature Review 7](#_Toc196682533)

[**3.1 Assessment of current existing quality measures and factors influencing the quality of post-acute care** 8](#_Toc196682534)

[**3.2 Factors influencing the selection of post-acute care** 8](#_Toc196682535)

[**3.3 Development of new tools for post-acute care selection** 9](#_Toc196682536)

[4. Data 9](#_Toc196682537)

[**4.1 Description of the data set** 10](#_Toc196682538)

[**4.1.1 Structure of the Dataset** 10](#_Toc196682539)

[**4.1.2 Statistical metrics for numerical variables** 11](#_Toc196682540)

[**4.2 Descriptive Statistics - Variable understanding** 12](#_Toc196682541)

[**4.3 Exploratory Data Analysis (EDA)** 17](#_Toc196682542)

[**4.3.1 Dealing with missing values** 21](#_Toc196682543)

[**4.3.2 Data exclusion** 22](#_Toc196682544)

[**4.3.3 Outlier Analysis** 23](#_Toc196682545)

[**4.3.4 Dimension Reduction** 26](#_Toc196682546)

[**1.** **PCA Analysis for Percent of beneficiaries identified with chronic conditions** 26](#_Toc196682547)

[**2.** **PCA Analysis for Percent of episodes with a primary diagnosis of disease** 27](#_Toc196682548)

[5. Methodologies 29](#_Toc196682549)

[**5.1 Principal Component Analysis (PCA)** 29](#_Toc196682550)

[**5.2 Exhaustive search for variable selection** 29](#_Toc196682551)

[**5.3 Multiple linear regression** 29](#_Toc196682552)

[**5. 4 Ridge / Lasso / Elastic Net Regression** 30](#_Toc196682553)

[6. Results 30](#_Toc196682554)

[**6.1 Multiple Linear Regression (for all service categories)** 30](#_Toc196682555)

[**6.2 Multiple Linear Regression (by each service category)** 31](#_Toc196682556)

[**6.2.1 HH Service Category** 31](#_Toc196682557)

[**6.2.2 HOS Service Category** 32](#_Toc196682558)

[**6.2.3 SNF Service Category** 32](#_Toc196682559)

[**6.2.4 LTC Service Category** 33](#_Toc196682560)

[**6.2.5 IRF Service Category** 33](#_Toc196682561)

[**6.3 Elastic Net Regression (for all records)** 34](#_Toc196682562)

[**6.4 Model comparison** 35](#_Toc196682563)

[**6.4.1 Multiple Linear Regression vs Elastic Net Regression (for all service categories)** 35](#_Toc196682564)

[**6.4.2 Multiple Linear Regression (for all service categories) vs Multiple Linear Regression (by each service category)** 35](#_Toc196682565)

[7. Conclusions 36](#_Toc196682566)

[Bibliography 37](#_Toc196682567)

## **List of Figures**

[Figure 1: Financial Flow of USA Healthcare System 6](#_Toc196682568)

[Figure 2 Projected Number of Older Adults and Children 6](#_Toc196682569)

[Figure 3: Count of Variables by Data Type 11](#_Toc196682570)

[Figure 4: Trends in Unique Providers by Service Category 12](#_Toc196682571)

[Figure 5: Count of Providers by State, Service Category 13](#_Toc196682572)

[Figure 6: Trends in Unique Beneficiaries by Service Category, National Level 13](#_Toc196682573)

[Figure 7: Count of Beneficiaries by State, Service Category 14](#_Toc196682574)

[Figure 8: Service Category and Total Medicare Payment Amount 14](#_Toc196682575)

[Figure 9: Average Medicare Payment Amount to the Provider by Service Category 15](#_Toc196682576)

[Figure 10: Average Medicare Payment Amount per Beneficiary by Service Category 15](#_Toc196682577)

[Figure 11: Distribution of Nursing Visits Per Beneficiary by Service Category 16](#_Toc196682578)

[Figure 12: Distribution of Physical Therapy Minutes Per Beneficiary by Service Category 16](#_Toc196682579)

[Figure 13: PAC Facility Usage by Age Group 17](#_Toc196682580)

[Figure 14: Correlation Heatmap for Providers’ Data 18](#_Toc196682581)

[Figure 15: Correlation Heatmap for Medicare Payment and Visits 19](#_Toc196682582)

[Figure 16: Nursing Visits (incl. Outliers) 19](#_Toc196682583)

[Figure 17: Nursing Visits (excl. Outliers) 20](#_Toc196682584)

[Figure 18: Social Work Visits (incl. Outliers) 20](#_Toc196682585)

[Figure 19: Social Work Visits (excl. Outliers) 20](#_Toc196682586)

[Figure 20: Aide Visits (incl. Outliers) 21](#_Toc196682587)

[Figure 21: Aide Visits (excl. Outliers) 21](#_Toc196682588)

[Figure 22: Boxplot for Distinct Beneficiary Count 23](#_Toc196682589)

[Figure 23: Episode or Stay Count 24](#_Toc196682590)

[Figure 24: Days of Service 24](#_Toc196682591)

[Figure 25: Total Charge Amount 25](#_Toc196682592)

[Figure 26: Total Amount Allowed by Medicare 25](#_Toc196682593)

[Figure 27: Total Medicare Payment Amount 25](#_Toc196682594)

[Figure 28: Total Medicare Standard Payment 26](#_Toc196682595)

[Figure 29: Scree Plot of PCA for Percent of Beneficiaries Identified with Chronic Condition 27](#_Toc196682596)

[Figure 30: Contribution of Variables to PC1 Chronic Conditions 27](#_Toc196682597)

[Figure 31: Scree Plot of PCA for Percent of Episodes with a Primary Diagnosis of Disease 28](#_Toc196682598)

[Figure 32: Contribution of Variables to PC1 for Primary Diagnosis 29](#_Toc196682599)

## **List of Tables**

[Table 1: Important Variables for Data Analysis 10](#_Toc196683947)

[Table 2: Structure of the Dataset 11](#_Toc196683948)

[Table 3: Statistical Metrics for the Numeric Variables 12](#_Toc196683949)

[Table 4: Variable Details Considered for Heatmaps 17](#_Toc196683950)

[Table 5: Variable Exclusion 23](#_Toc196683951)

[Table 6: Importance of components for Chronic Condition Variables 26](#_Toc196683952)

[Table 7: Importance of Components for Primary Diagnosis Variables 28](#_Toc196683953)

[Table 8: Coefficients of Multiple Linear Regression Model for all service categories 30](#_Toc196683954)

[Table 9: Prediction Performance of Multiple Linear Regression Model for all service categories 31](#_Toc196683955)

[Table 10: Coefficients of Multiple Linear Regression Model for HH 31](#_Toc196683956)

[Table 11: Prediction Performance of Multiple Linear Regression Model for HH Records 31](#_Toc196683957)

[Table 12: Coefficients of Multiple Linear Regression Model for HOS 32](#_Toc196683958)

[Table 13: Prediction Performance of Multiple Linear Regression Model for HOS Records 32](#_Toc196683959)

[Table 14: Coefficients of Multiple Linear Regression Model for SNF Records 32](#_Toc196683960)

[Table 15: Prediction Performance of Multiple Linear Regression Model for SNF Records 33](#_Toc196683961)

[Table 16: Coefficients of Multiple Linear Regression Model for LTC Records 33](#_Toc196683962)

[Table 17: Prediction Performance of Multiple Linear Regression Model for LTC Records 33](#_Toc196683963)

[Table 18: Coefficients of Multiple Linear Regression Model for IRF Records 34](#_Toc196683964)

[Table 19: Prediction Performance of Multiple Linear Regression Model for IRF Records 34](#_Toc196683965)

[Table 20: Coefficients of Elastic Net Regression 34](#_Toc196683966)

[Table 21: Prediction Performance of Elastic Net Regression 35](#_Toc196683967)

[Table 22: Performance Comparison between Multiple Linear and Elastic Net Regression 35](#_Toc196683968)

[Table 23: Performance comparison of all service categories with each service category 36](#_Toc196683969)

## **1. Abstract**

Post-acute care (PAC) facility utilization cost is an essential driver for Medicare payments yet understanding the factors that influence the total payment remains limited. This study applies predictive analytics methodologies to identify and quantify the impact of facility service metrics on Medicare payments. Using a provider-level service data set with added principal component features, linear regression models were trained and evaluated.

Results suggested that the combination of the number of beneficiaries utilizing the facility, total service days, and occupational therapy minutes have a positive influence on the payment, while nursing visits, social work visits, physical therapy minutes have a negative impact on the payment. The findings have proven that the greater the number of patients utilizing the PAC facilities and the greater number of days they get the services, the Medicare payment increases. While occupational therapy minutes have a positive impact on the Medicare payment, the Physical therapy minutes have a negative impact.

**Keywords:** CMS Data, Post Acute Care, PAC, HH, Total Medicare Cost, Multiple Linear Regression, Elastic Net Search, Home Health, Hospice, Skilled Nursing Facilities, Long Term Care Facilities, Inpatient Rehabilitation Facility.

## **2. Introduction**

The U.S. healthcare system is intricate and complex, involving many coverages, providers, and payment models [1]. One of the most demanding aspects of this system is the need for post-acute and hospice care, especially for elderly patients requiring [medical rehabilitation services](https://www.gminsights.com/industry-analysis/medical-rehabilitation-services-market), long-term care, and specialized recovery solution [2]. Post-acute care includes services provided in skilled nursing facilities (SNFs), inpatient rehabilitation facilities (IRFs), long-term care hospitals (LTCHs), and home health agencies [3]. Hospice care, on the other hand, focuses on end-of-life support for terminally ill patients [2], [3], [4]. The complexities of accessing these services arise from fragmented healthcare structures, inconsistent quality standards, and financial constraints that burden patients and their families [1]- [5].

The healthcare system in the US does not provide universal coverage for everyone. It is a combined system, where publicly financed government Medicare and Medicaid coverage exists with private health insurance providers. Out-of-pocket payments and insurance coverage provide most of the financial coverage for healthcare. The following Figure 1 provides an overview of the money flow of the US healthcare system [1]. This clearly illustrates that the flow of funds is not straightforward, providers are paid by many sources including State and Federal governments, which often mandate traceability and regulations.

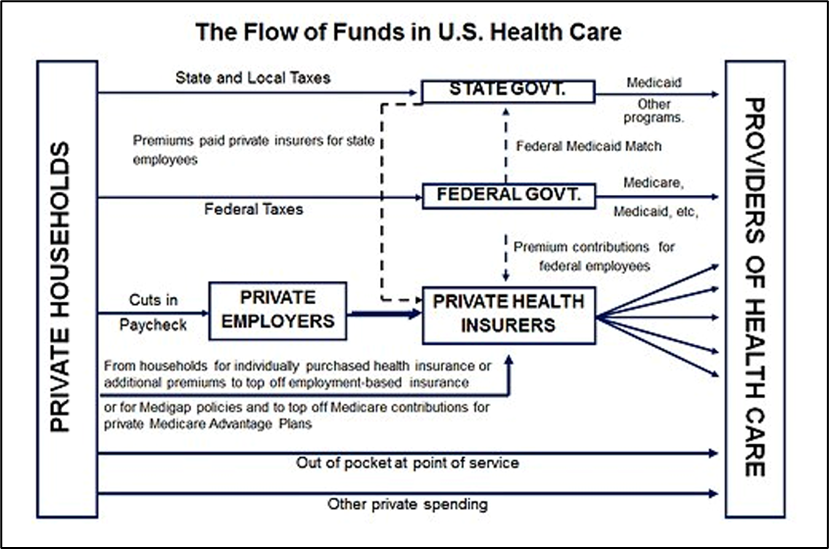


Figure 1: Financial Flow of USA Healthcare System

Source: Reinhardt UE. The Money Flow from Household to Health Care Providers (2011)

There is an importance in addressing these challenges. As the U.S. population continues to age as illustrated in Figure 2, with projections indicating that nearly 20% of Americans will be 65 or older by 2030, [6], the demand for post-acute and hospice care is expected to rise significantly. Many individuals find it hard to get the best care facilities due to a lack of information, inconsistency in providing metrics, and financial limitations. Though Medicare is the primary insurance provider for elderly individuals covering both post-acute and hospice care yet there are gaps in terms of affordability and accessibility to services [5]. Identifying the severity of the quality, cost variations and the distribution of care facilities plays a crucial role in enhancing data-driven decisions and improving the patient outcomes.

A graph of ageing and ageing

AI-generated content may be incorrect.

Figure 2 Projected Number of Older Adults and Children

Source: <https://x.com/uscensusbureau/status/973926724776939520?mx=2>

Elderly patients face many challenges getting care through Medicare, it starts with the complex enrollment process, high out-of-pocket costs [7], limited coverage for prescriptions, difficulty finding healthcare providers who accept Medicare, limited or no coverage for long-term post-acute care, cost and coverage gap for various medical services, complexity in the coordination of benefits, and lack of accessible assistance and support [8]. All these challenges contribute to a lot of distress while they are already going through a difficult medical situation and a lack of insight into the data to choose the right facility adds to these existing challenges.

Decisions about what includes appropriate medical care are very complex and necessary for aging populations. Patients have their own beliefs and choices. Physicians treating patients might have a different perspective which may or may not align with what the patient wants [9]. Knowledge and availability of available interventions is limited. Shortage of acute care facilities, complex Medicare and limitations in insurance covering eligibility requirements, high costs and financial struggles, lack of reliable transportation, and disparities in accessing health care facilities [10] are some of the challenges faced by patients while choosing the PAC (Post Acute Care) facilities. Choosing the right setting for a PAC has major challenges. Cost & Location are important factors in choosing. Current available public data from CMS (Centers for Medicare and Medicaid services) “Five-star rating system” and “Choosing Wisely Recommendations” measures quality of facilities and is complicated, their validity and fairness is questioned [11].

The elderly face many hurdles in getting acute and post-acute care as they must navigate through complex healthcare systems, unnecessary physician dependence, and often get confused by Medicare information. Although there are quality measures available through CMS, they are not user-friendly and are bulky, thus it is hard for elderly patients to make sound decisions. Health literacy and administrative wastefulness, such as previous authorizations, further complicate access to on-time care. The resolution of these problems demands better healthcare information, enhanced patient-clinician communication, and streamlined administrative procedures to improve decision-making and access to care.

*This research aims to identify the PAC facility attributes that contribute to the Medicare payment amount, for all service categories and for each of the service categories. Identifies the state that has a greater number of providers and is serving more patients.*

To provide a better understanding on the quality, cost, and accessibility of post-acute and hospice care, this research uses a combination of statistical and machine learning methodologies to *help individuals, family members, and caretakers leveraging the data to help with decision-making based on their needs*. The research leverages predictive analytics (regression analysis, decision trees, random forests and clustering techniques), EDA (Exploratory Data Analysis), Geospatial analysis, and machine learning methods to identify the patterns in healthcare facilities and to derive the key performance indicators affecting patient outcomes based on which elderly can select the facilities that fit their needs.

The paper is organized as follows. Section 3 reviews existing literature on similar topics. Section 4 provides data insights from the exploratory data analysis, to understand the data type, outliers, value ranges and missing values. Section 5 outlines the methodologies used to understand the correlation between variables, dimension reduction and brief information on the models being used for this analysis. Section 6 focuses on the results and comparison between various models used. Finally, Section 7 summarizes the findings, limitations of the data set and current research.

## **3. Literature Review**

The selection of post-acute and hospice care is broadly studied in literature, primarily discussing measuring the quality of care when selecting. Several kinds of literature identify factors that influence quality. In addition to the quality, various factors affecting patients’ and family's decision-making are also discussed and identified. Finally, supporting tools for better selection are developed through several pieces of literature. We explain and contrast our research with them in the sections below.

### **3.1 Assessment of current existing quality measures and factors influencing the quality of post-acute care**

There is a lot of research studied to evaluate the quality measures of post-acute care, as it allows users to compare facility performance and making choices [12]. Three articles assess the validity of currently available quality measures that are used by patients, their families, and caregivers to evaluate care services. Castle and Ferguson [13] examine the advantages and disadvantages of the two quality indicators, facility quality indicator profile report and nursing home compare. They conclude that those indicators are not perfect for differentiating the real quality of each facility. Olenski and Sacher [14] conduct a similar study as Castle and Ferguson [13] but using analytical models to check correlation between public report card system provided by nursing home compare and quality. The authors find that the conventional measurement has almost no correlation with their survival-based quality estimates and, thus, it is less useful. Evaluations have been carried out on the CMS star rating system and find that this quality indicator does not reflect consumer satisfaction with the facility reported and recommends improvement by reflecting consumer perspectives [15]. The above studies suggest the defect of performance of those measurements and necessity of the more effective and convenient data, but do not propose any alternative quality measures to support selection. Also, the scope is limited to nursing homes.

There are papers studying how much current quality measures affect user’s choice of care. [16] shows the impact of the introduction of two sets of ratings, quality of care, and patient experience, on consumer's home health agency (HHA) selection. The findings highlight that both ratings positively increase the user’s likelihood of choosing highly quality and highly satisfactory facilities. It suggests that the rating system somewhat works but is limited to HHA. [17] also study the correlation of publicly available scores named report card scores and patient choice for nursing home to confirm its reliability. By taking a market fixed-effect approach, it reveals that the correlation is positive and suggests that patients who refer to the scores tended to select facilities with higher post-acute care quality. However, the importance or accuracy of these quality measures is still being questioned.

Overall, the above literature suggests that consumers can refer to current existing quality indicators when selection, but linkage with actual quality is weak or limited to some types of care facilities. Furthermore, the variable to choose the best possible care is only limited to quality. In contrast, our research study attempts to identify key performance indicators which were unexploited before, by validating a set of variables available in CMS data.

Several research papers identify factors that may influence the quality of post-acute care facilities. One research highlights the problem of hospital readmission from PAC facilities and necessity to reduce the readmission rates to improve quality of care. It conducts risk factor analysis and finds that frequent PAC physicians' visits, longer length of stay are the factors to increase readmission rate, which can be considered to lower the quality of the facilities [18]. Another research, [19], studies the impact of rehabilitation therapies on the quality of care and desired outcome of community discharge using the retrospective study of datasets. It finds that the intensity of the therapy in SNFs minimizes length of stay and leads to early return to community. These studies implicate some variables which can be used to evaluate each care facility but do not suggest practical tools to improve user’s decision making.

### **3.2 Factors influencing the selection of post-acute care**

A series of research papers are published outlining the factors influencing the selection of post-acute care. In research on information collected from interviews to estimate the factors that influence the choice of a post-acute care services, five main factors are identified, medical professionals’ suggestions, health care accessibility, continuity and coordination of care, preferences and experience of patients and their relatives and friends, and economic factors [20]. These factors may assist patients and family members’ decision-making. Two other similar studies were conducted but focusing on only SNFs. One of them involved conducting interviews to identify what factors are important in decision-making and revealed that most people select a facility based on its location, due to lack of comprehensive information [21]. The other one uses qualitative methods to identify factors affecting patients’ and caregivers' decision-making about utilizing SNFs and suggests that a lack of choices and tools to evaluate best possible care except medical provider’s recommendations lead to dissatisfaction of their decision [22]. Another research mainly argues about the cost perspective [23]. The authors highlight the need to ensure adequate long-term care services at an affordable cost by explaining how the current long-term care system is problematic and a burden to patients. These papers are valuable for understanding external factors but do not suggest concrete ways or tools for patients to choose the best possible care among available facilities.

### **3.3 Development of new tools for post-acute care selection**

Several studies focus on the development of novice tools that help users to choose the best post-acute care. Since limited indicators are available as of early 2000, [4] develops home care quality indicators (HCQIs) by involving researchers, clinicians and policy makers from three countries and examining multiple literatures. It aims to be used by various users such as consumers to support evidence-based decision-making about the quality of home care services. The HCQIs are new tools contributing to quality improvement for home care. [24] concerns the quality of care of nursing homes. It develops a new quality indicator named MDS Quality Indicators through clinical inputs, empirical analysis and field-testing to evaluate nursing home quality. It suggests that the indicator could be a useful tool for nursing home residents, families, and caregivers to improve their ability to make more sensible decisions about selection. To evaluate the quality of care and facility performance, [12] establishes a new quality measure of IRF, named discharge self-care functional status by conducting cohort sample study including patients from 38 IRFs. The research proved that it has strong ability and reliability to compare across IRF facilities. Lastly, the other research emphasizes the need to enhance the understanding of broad aspects of LTC services to support an aging society. It explains the several statistical forecasting models that can predict the need, demand, and cost of the long-term care in Asian and Pacific regions by reviewing microsimulation or macrosimulation models. Such models can grasp the general needs for elderly people who need care [25]. In contrast, our study aims to develop more holistic tools that support patients’ decision-making not only by indicating quality of the service but also by including cost, location, and so forth. Furthermore, the scope of the service is broader, including SNFs, IRFs, LTCHs, HHA, and Hospice care.

In conclusion, existing literature primarily focuses on quality as a variable of measuring performance of the PAC facilities. Several papers suggest other decision-making factors such as the medical provider’s recommendation, facility’s location, costs but there are no specific tools to compare these indicators among facilities. In addition, most of the study targets just one or two PAC settings, while our paper aims to cover all categories of services. To fill the gaps, our research contributes to the literature by applying machine learning models and predictive analytics to predict the cost for each PAC settings and provide data driven decision making tools for patients and families who struggle to make choices of PAC services that meet their demand.

## 4. **Data**

The dataset contains around 89,393 rows and 88 columns, including provider-level, state-level, and national-level data for various post-acute care (PAC) settings, covering home health, hospice, SNFs, IRFs, and LTCHs. Includes details on beneficiary demographics, service utilization, conditions, and Medicare payments. Key metrics like total stays, charges, payments, and therapy minutes are also included. Additionally, it tracks diagnoses, dual eligibility, and rural classifications for the beneficiaries.

[Medicare Post-Acute Care and Hospice - by Geography & Provider](https://data.cms.gov/search?keywords=Medicare%20Post-Acute%20Care%20and%20Hospice%20-%20by%20Geography%20%26%20Provider%20&sort=Relevancy)

The following table has important variables that are being considered for the analysis.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Label** | **Description** |
| SRVC\_CTGRY | Service Category | Identifies the PAC setting: HH, hospice, SNFs, IRFs, or LTCHs. |
| BENE\_DSTNCT\_CNT | Distinct Beneficiaries | Number of unique Medicare beneficiaries with at least one paid claim in the calendar or fiscal year. |
| TOT\_EPSD\_STAY\_CNT | Episode or Stay Count | For home health, this is the total count of 60-day episodes in CY. For hospice, SNF, IRF, and LTCH this is the total count of stays provided in the fiscal year. |
| TOT\_SRVC\_DAYS | Days of Service | Total count of covered days delivered by a provider in the calendar or fiscal year. |
| TOT\_CHRG\_AMT | Total Charge Amount | Total charges submitted by the provider. |
| TOT\_ALOWD\_AMT | Total Allowed Amount | Total of the Medicare allowed amount; this figure is the sum of the amount Medicare pays, the deductible and coinsurance amount that the beneficiary is responsible for paying, and any amount that a third party is responsible for paying. This applies only to the SNF, IRF, and LTCH settings. |
| TOT\_MDCR\_PYMT\_AMT | Total Medicare Payment Amount | The total amount that Medicare paid after deductible and coinsurance amounts have been deducted. |

Table 1: Important Variables for Data Analysis

### **4.1 Description of the data set**

Descriptive analysis is the primary step in understanding a dataset. This analysis helps to explore characteristics of variables and summarize the data to derive meaningful insights into distributions, trends, identifying and working with potential anomalies. In this section we will specifically, we will focus on the following aspects.

* Understanding the structure and types of variables
* Computing summary statistics: numerical measures of center and spread. A few of them include calculating mean, median, min, max, range, variance, standard deviation and so on.
* Visualization for key variables
* Analyzing potential correlation and related trends between various variables

#### **4.1.1 Structure of the Dataset**

The dataset dimensions include 88 columns and 89,392 rows of data.

Table 2 provides the top 10 details of the filtered dataset to understand the structure of dataset i.e. identifying variable names and datatypes. Count of the variable types i.e. number of categorical, numerical and logical data types exist in the dataset. Identifying number of unique values which helps to detect inconsistencies such as unexpected categories (e.g., typos like 'Mle' instead of 'Male') or out-of-range numerical values (e.g., negative service days or unusually high payment amounts), which may indicate data entry errors or anomalies.". Checking for any missing values helps to understand the incompleteness of the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Variable Description | Data\_Type | Unique\_Values | Missing\_Values |
| YEAR | YEAR | INTEGER | 3 | 0 |
| YEAR\_TYPE | YEAR TYPE | CHARACTER | 2 | 0 |
| SMRY\_CTGRY | SUMMARY CATEGORY | CHARACTER | 3 | 0 |
| SRVC\_CTGRY | SERVICE CATEGORY | CHARACTER | 5 | 0 |
| PRVDR\_ID | PROVIDER ID | CHARACTER | 31641 | 0 |
| PRVDR\_NAME | PROVIDER NAME | CHARACTER | 32145 | 0 |
| PRVDR\_CITY | PROVIDER CITY | CHARACTER | 5733 | 0 |
| STATE | STATE | CHARACTER | 53 | 0 |
| PRVDR\_ZIP | PROVIDER ZIP CODE | CHARACTER | 10697 | 0 |
| BENE\_DSTNCT\_CNT | DISTINCT BENEFICIARIES | INTEGER | 3277 | 0 |

Table 2: Structure of the Dataset

Figure 3 shows the count of variables by data types. The dataset contains most variables falling under character datatype.

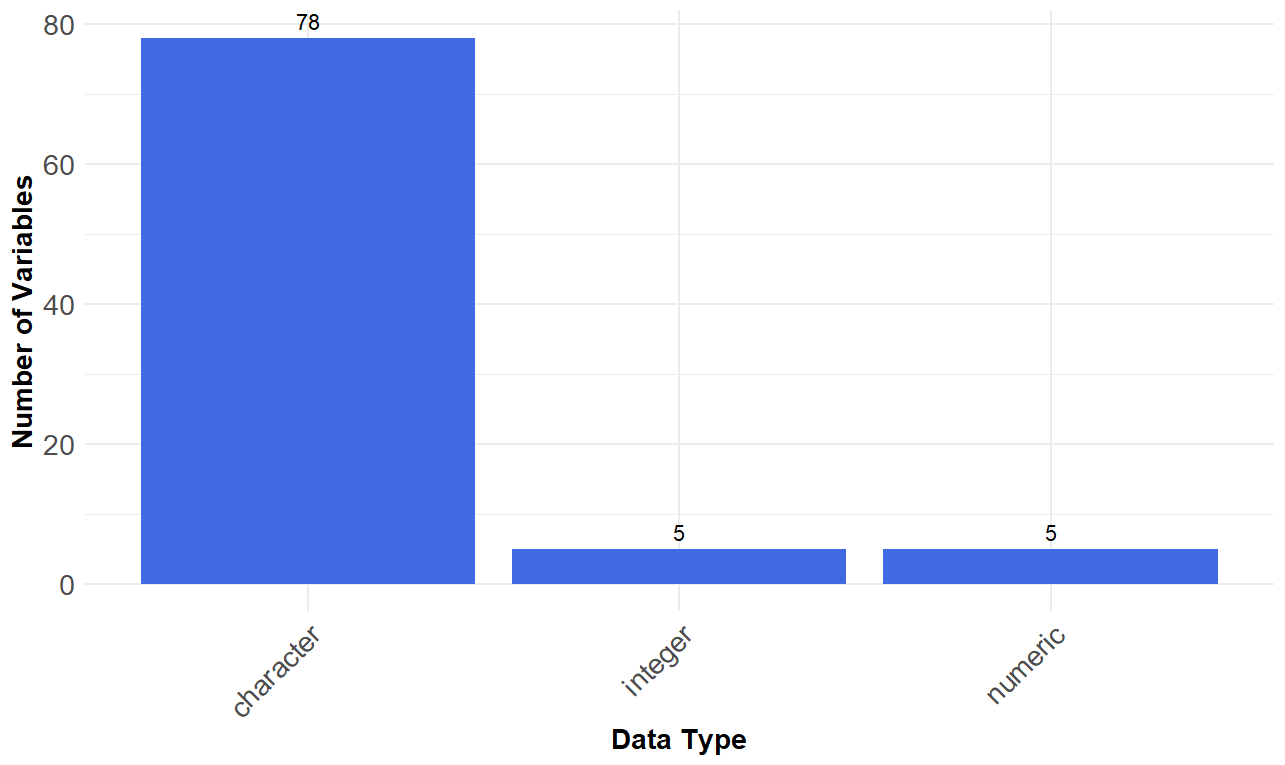


Figure 3: Count of Variables by Data Type

#### **4.1.2 Statistical metrics for numerical variables**

Table 3 shows the statistical metrics for the numeric variables.

* Cost and service variables are strongly skewed, which is indicated by wide differences between mean and median and extremely high standard deviations (e.g., total charge standard deviation is more than 36 million). This suggests extreme outliers and non-uniform distribution of service volumes or prices across providers.
* Demographic variables are of low variability, as reflected in small standard deviations (e.g., mean age SD ≈ 4.5) and restricted ranges. This suggests a stable beneficiary profile across the dataset, reducing the potential for demographic bias in later analysis.
* Provider-level distinctions are robust, as indicated by large value ranges (e.g., different beneficiary count ranges from 11 to more than 3 million). Such robust variation in provider capacity and service size necessitates segmentation in relative evaluations.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable name** | **Variable Description** | **Mean** | **Median** | **Min** | **Max** | **Range** | **Variance** | **Std\_Dev** |
| BENE\_DSTNCT\_CNT | DISTINCT BENEFICIARIES | 6.63E+02 | 107 | 11 | 3.02E+06 | 3.02E+06 | 4.70E+08 | 2.17E+04 |
| TOT\_EPSD\_STAY\_CNT | EPISODE OR STAY COUNT | 1.30E+03 | 155 | 11 | 9.27E+06 | 9.27E+06 | 2.96E+09 | 5.44E+04 |
| TOT\_SRVC\_DAYS | DAYS OF SERVICE | 2.53E+04 | 3494 | 19 | 1.29E+08 | 1.29E+08 | 8.20E+11 | 9.05E+05 |
| TOT\_CHRG\_AMT | TOTAL CHARGE AMOUNT | 1.27E+07 | 1612532 | 4180 | 3.68E+10 | 3.68E+10 | 1.26E+17 | 3.54E+08 |
| TOT\_ALOWD\_AMT | TOTAL ALLOWED AMOUNT | 8.28E+06 | 1486563 | 3041 | 3.18E+10 | 3.18E+10 | 6.52E+16 | 2.55E+08 |
| TOT\_MDCR\_PYMT\_AMT | TOTAL MEDICARE PAYMENT AMOUNT | 7.73E+06 | 1324920 | 3041 | 2.65E+10 | 2.65E+10 | 5.47E+16 | 2.34E+08 |
| TOT\_MDCR\_STDZD\_PYMT\_AMT | TOTAL MEDICARE STANDARD PAYMENT AMOUNT | 7.74E+06 | 1391369 | 3022 | 2.65E+10 | 2.65E+10 | 5.48E+16 | 2.34E+08 |
| BENE\_AVG\_AGE | BENEFICIARY AVERAGE AGE | 7.86E+01 | 79 | 46 | 9.30E+01 | 4.70E+01 | 2.04E+01 | 4.51E+00 |
| BENE\_AVG\_RISK\_SCRE | AVERAGE RISK SCORE | 2.54E+00 | 2.43 | 0.67 | 9.76E+00 | 9.09E+00 | 4.01E-01 | 6.33E-01 |

Table 3: Statistical Metrics for the Numeric Variables

### **4.2 Descriptive Statistics - Variable understanding**

In this section, key variables that are relevant to our research are explained with visualization.

Below is the abbreviation of each PAC services used in figures:

* Skilled Nursing Facilities - SNF​
* Long Term Care - LTC​
* Inpatient Rehabilitation Facilities - IRF​
* Home Health Agencies- HH
* Hospice Care - HOS

Figure 4 shows the count of providers by state and service category in 2022. It illustrates which state has more care facilities by each category. According to the chart, SNF and home health have the highest numbers of providers among all PAC settings. In terms of state as shown in Figure 5, provider’s numbers are dependent on population. Particularly, CA, TX, and FL are seen as the largest segments in each service category.

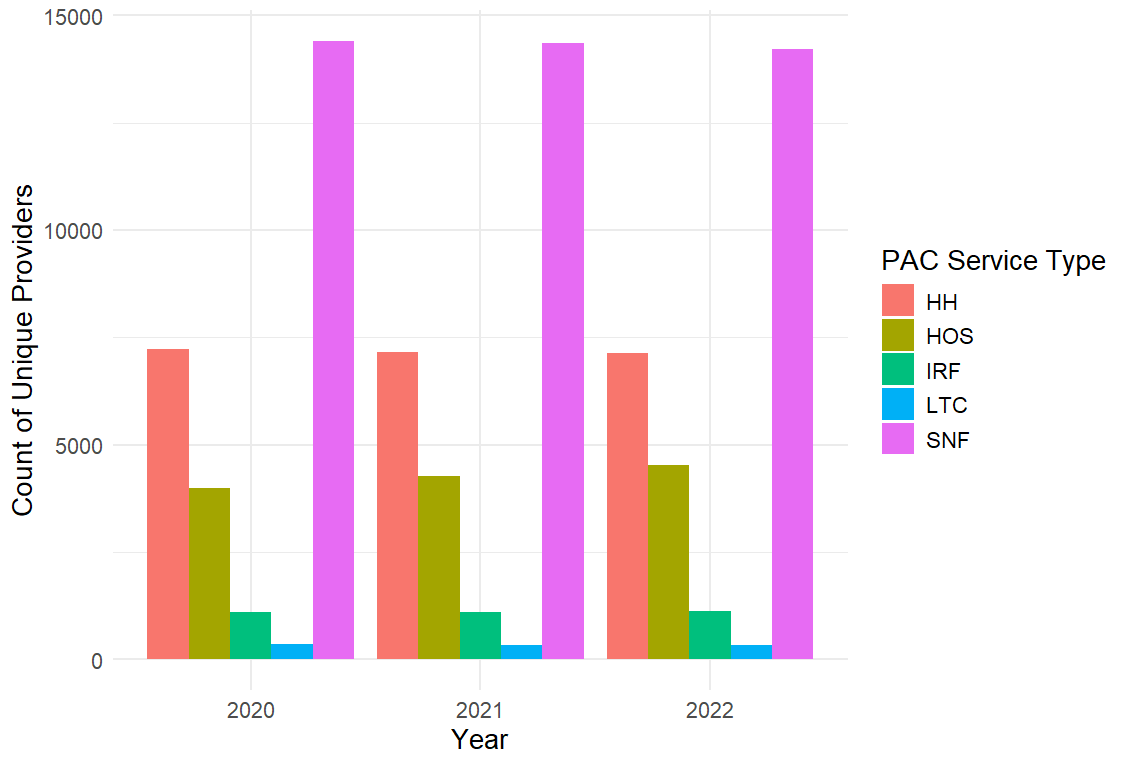


Figure 4: Trends in Unique Providers by Service Category

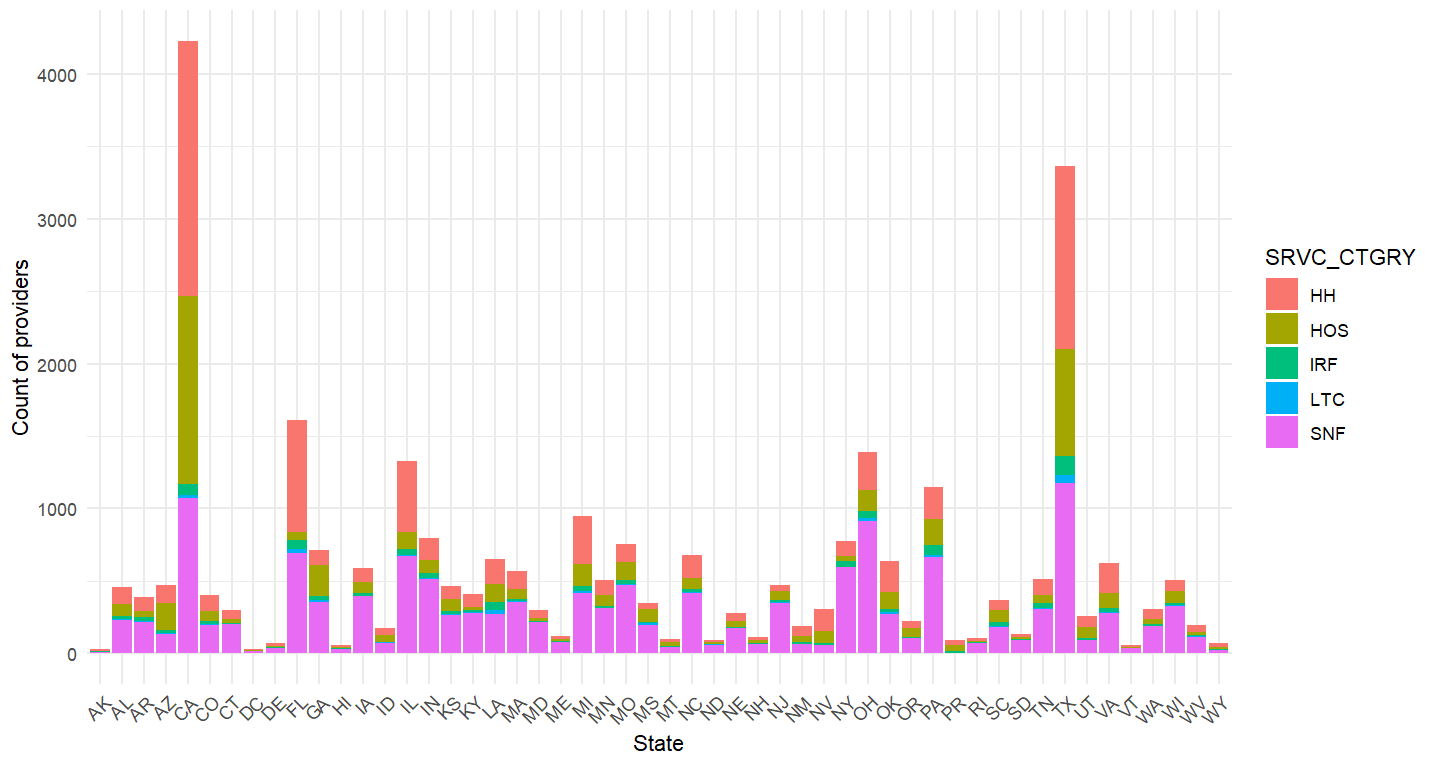


Figure 5: Count of Providers by State, Service Category

Figure 6 shows the trends in unique beneficiaries utilizing PAC services by service category and unique beneficiaries of all three years, i.e. 2020 to 2022 filtering at the national level. It illustrates which post-acute care facilities were used the most by unique beneficiaries over three years. According to the chart, HH (Home Health) is the highest post-acute care facility utilized and LTC (Long Term Care) is the least post-acute care facility utilized by unique beneficiaries over all three years.

Figure 7 shows the count of beneficiaries by state and service category in 2022. It illustrates which state serves the greatest number of patients by each category. As well as the count of providers, beneficiary numbers are also according to the state’s population. Particularly, CA, FL, TX, and NY are seen as the largest segments.

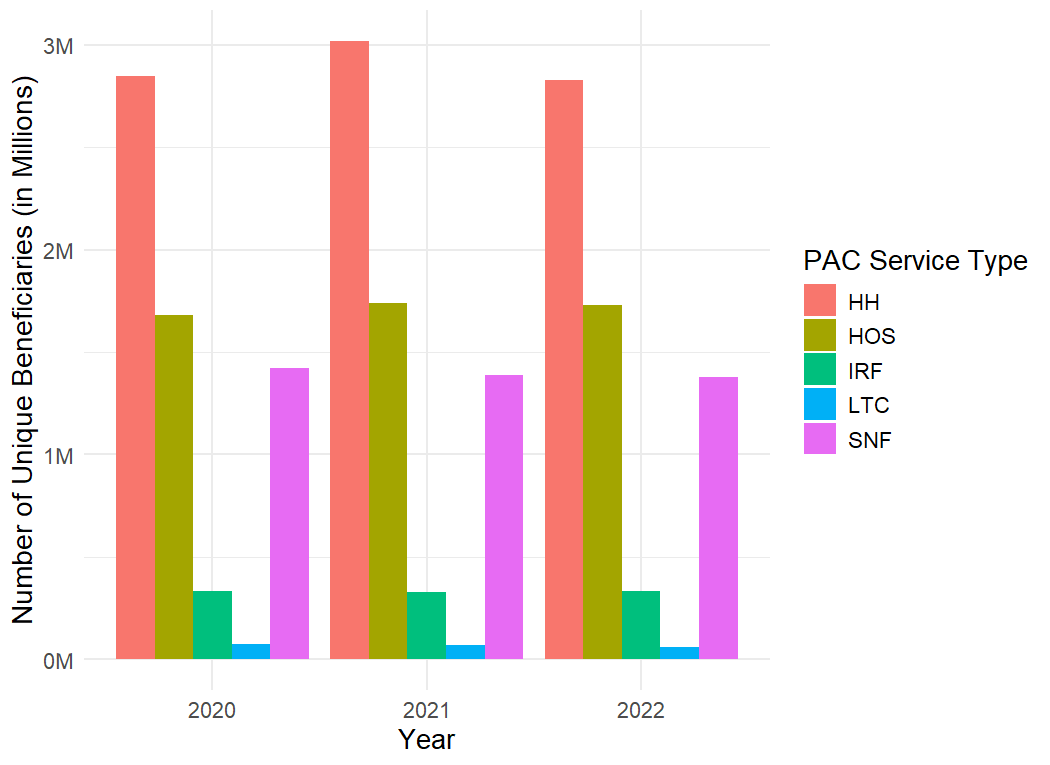


Figure 6: Trends in Unique Beneficiaries by Service Category, National Level

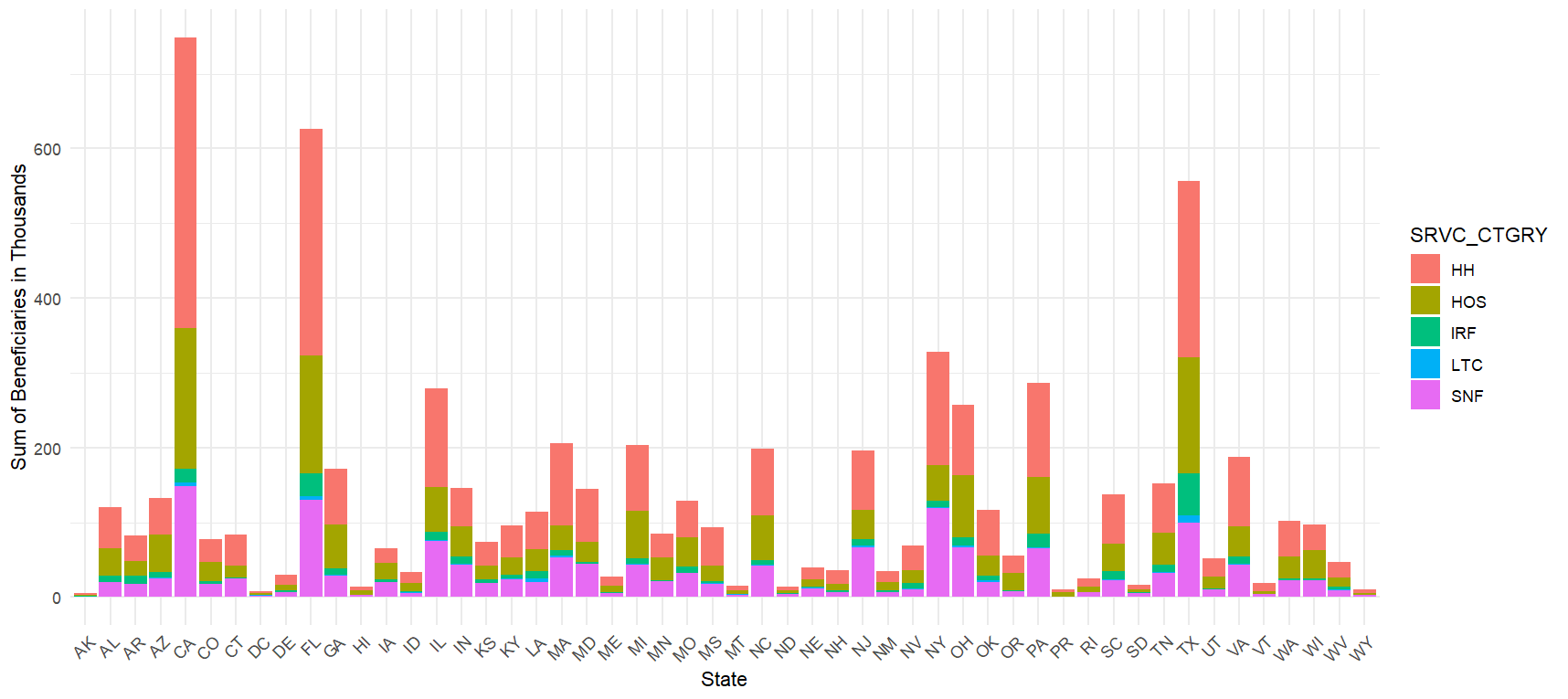


Figure 7: Count of Beneficiaries by State, Service Category

The sum of the total Medicare payment amount per service category for the three years. Figure 8 demonstrates that the SNF facility has the highest total payment for each of the years when we consider the sum of the amounts of all providers in the nation.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 8: Service Category and Total Medicare Payment Amount

Figure 9 illustrates the annual average cost that Medicare pays to each provider. The average payment to providers in the LTC and IRF categories is the highest among all categories. Figure 10 shows the average payment amount per person (beneficiary) in 2022. It explains how much money Medicare pays per person annually in each category. According to the chart, LTC is paid the most. The reason behind the different results from 3) is that SNF has more than 1 million beneficiaries while LTC has less than 50,000. The number of beneficiaries boosts the sum of the payment for SNF. However, when we look at the average cost per beneficiary, payment for LTC is much higher.

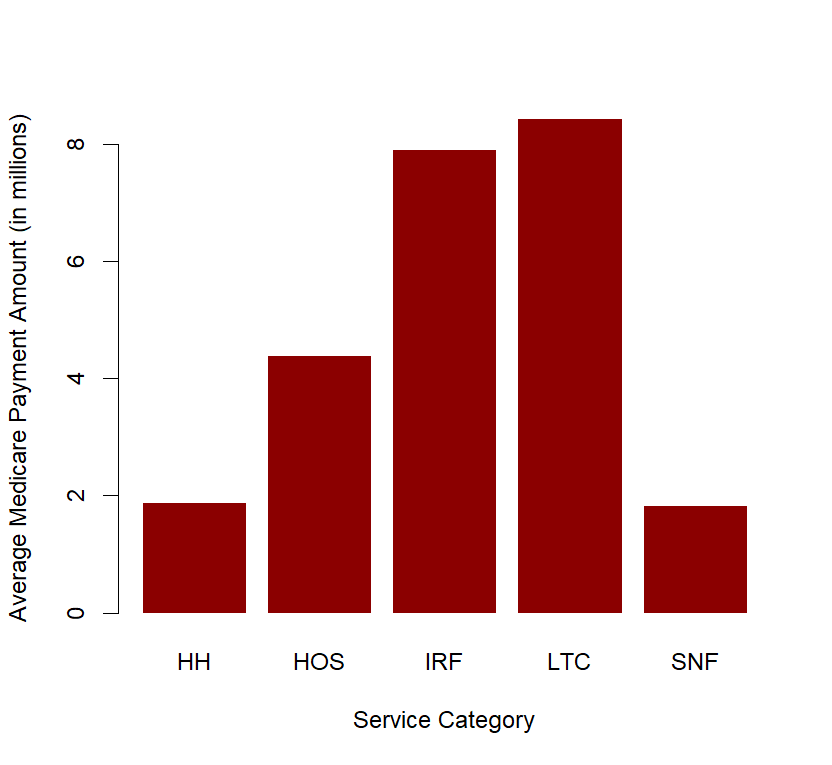


Figure 9: Average Medicare Payment Amount to the Provider by Service Category

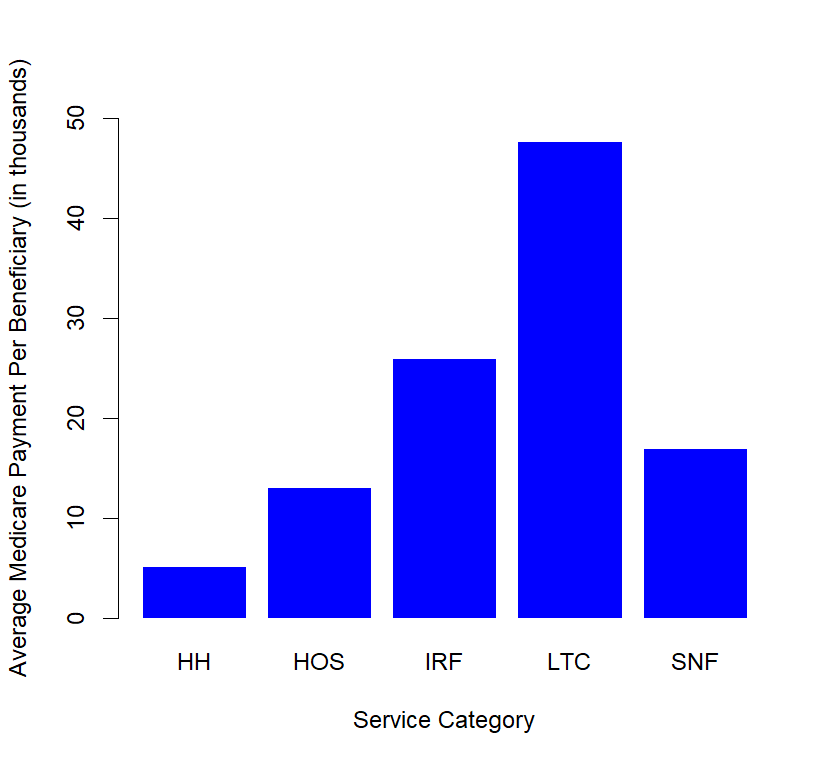


Figure 10: Average Medicare Payment Amount per Beneficiary by Service Category

Figure 11 shows the box plot of nursing visits per beneficiary by service category in 2022. It illustrates how frequently skilled nursing visits are provided by home health and hospice. Each observation is equal to the average nursing visits per person in each provider. In most of the observations, nursing visits are clustered around 10 visits for home health, and 20 visits for hospice per person (beneficiary). It is noteworthy that there are many outliers for both categories. It implies that the providers in outliers far too frequently visit patients for some specific reason. Although we will not be able to clarify the true reason from the data, our understanding is that too many nursing visits might lower the quality of care.

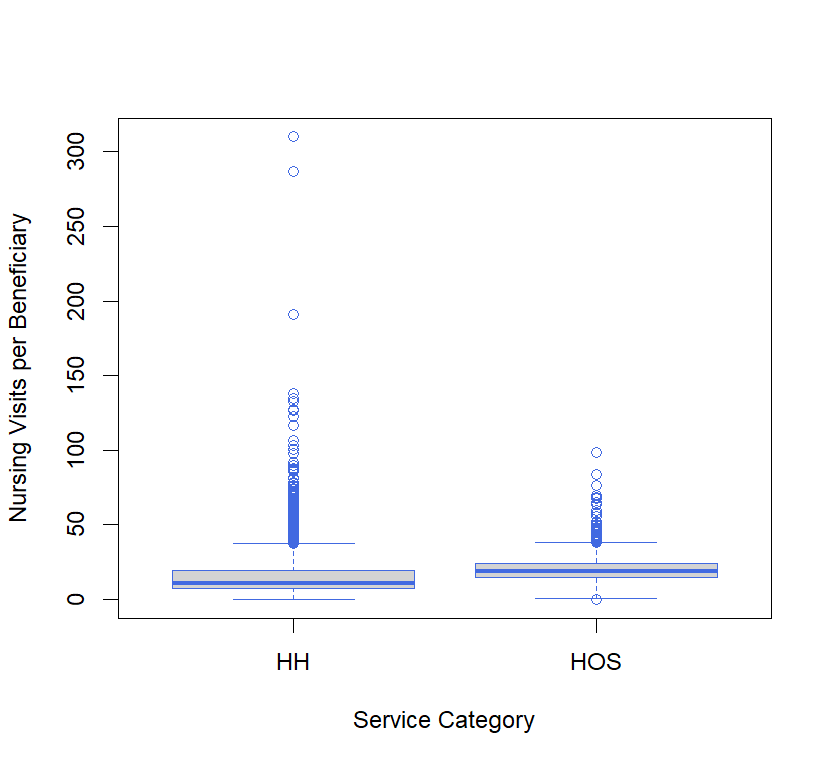


Figure 11: Distribution of Nursing Visits Per Beneficiary by Service Category

Figure 12 shows the boxplot of physical therapy minutes per beneficiary by service category in 2022. It illustrates how intense physical therapy is conducted per person (beneficiary) in home health and IRF. Each observation is equal to each provider. In most of the observations, therapy minutes are clustered around 400 minutes for home health, and 700 minutes for IRF per person (beneficiary). It is worth mentioning that home health has many outliers beyond the upper quartile, while IRF has one on the lower side. According to our observation of the data, this implies that the providers implementing longer therapy might give high quality service, while shorter therapy might mean insufficient quality.

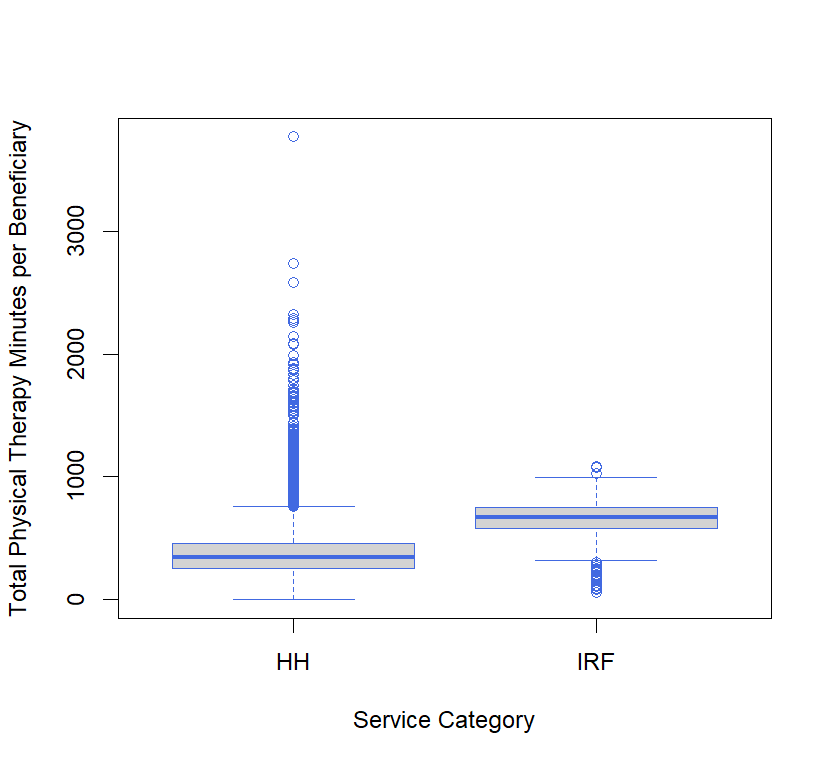


Figure 12: Distribution of Physical Therapy Minutes Per Beneficiary by Service Category

Figure 13 demonstrates the PAC Facility usage by age group. Data categorized by the average beneficiary age to understand the age group that is utilizing the PAC facilities the most, it has been observed that the 76-80 age group members are utilizing the most PAC facilities.

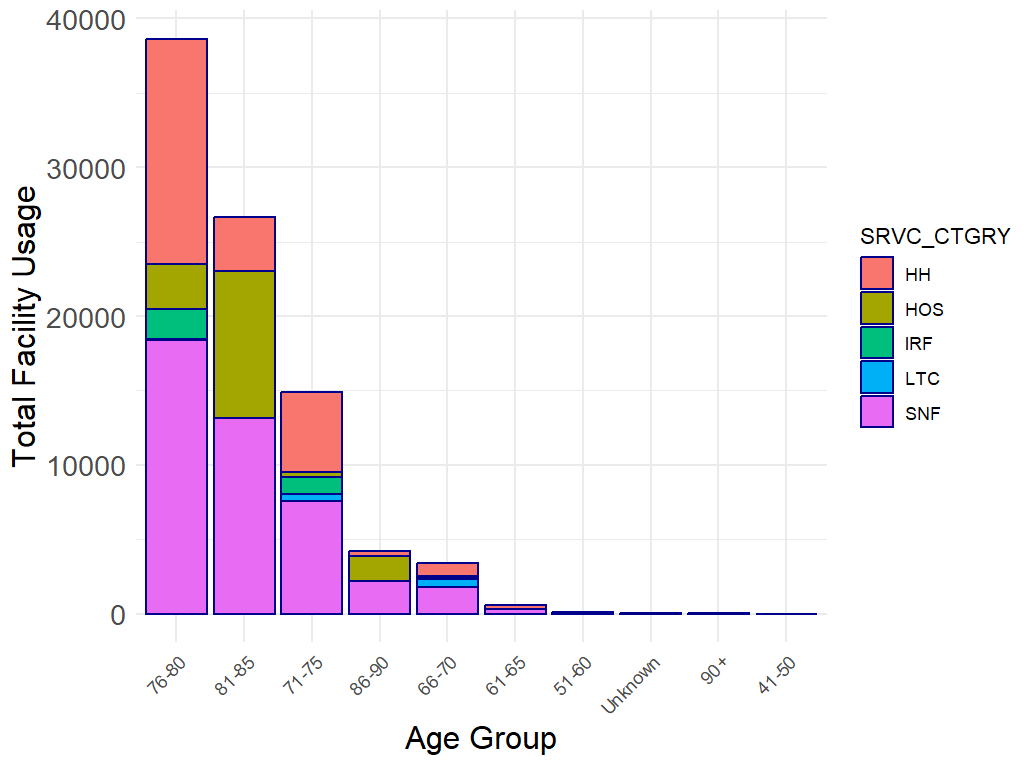


Figure 13: PAC Facility Usage by Age Group

### **4.3 Exploratory Data Analysis (EDA)**

Exploratory data analysis (EDA) is an important data analysis process that focuses on understanding the quality of variables, identifying trends, detecting anomalies, and identifying underlying patterns within the dataset using visual and quantitative methods. Through the EDA process, the data is cleaned, replaced with meaningful values where applicable, and made suitable for further modeling.

In this section, variables are analyzed to find key trends, relationships, and anomalies. Heatmaps and scatter plots are used to visualize the correlation between the variables. Scatterplots and box plots are used to detect outliers. Missing values are addressed to ensure the quality of data to support meaningful analysis. Principal Component Analysis is used for dimension reduction, i.e., analysis of reducing the number of variables.

A correlation analysis was done for the variables listed in Table 4 and Figure 14 has the correlation visualization.

|  |  |
| --- | --- |
| **Variable name** | **Variable Description** |
| BENE\_DSTNCT\_CNT | DISTINCT BENEFICIARIES |
| TOT\_EPSD\_STAY\_CNT | EPISODE OR STAY COUNT |
| TOT\_SRVC\_DAYS | DAYS OF SERVICE |
| TOT\_CHRG\_AMT | TOTAL CHARGE AMOUNT |
| TOT\_ALOWD\_AMT | TOTAL ALLOWED AMOUNT |
| TOT\_MDCR\_PYMT\_AMT | TOTAL MEDICARE PAYMENT AMOUNT |
| TOT\_MDCR\_STDZD\_PYMT\_AMT | TOTAL MEDICARE STANDARD PAYMENT AMOUNT |
| TOT\_OUTLIER\_PYMT\_AMT | TOTAL OUTLIER PAYMENT AMOUNT |
| BENE\_AVG\_AGE | BENEFICIARY AVERAGE AGE |
| BENE\_AVG\_RISK\_SCRE | AVERAGE RISK SCORE |

Table 4: Variable Details Considered for Heatmaps

Strong Positive Correlations

* The TOT\_SRVC\_DAYS variable which means the days of service has strong correlation with total allowed amount, charge amount, payment amounts, suggesting that which one tends to increase as service days are increases the amount related variables also increases.
* The total charge amount and total allowed amount show a very high positive correlation, indicating a direct relationship between billed charges and allowed amounts.

Moderate to High correlations

* The payment related variables total charge amount, total allowed amount, total Medicare payment amount and total Medicare standard amount show moderate positive correlation which implies higher amounts in one category may correlate with higher amounts in another.

Weak or Negative correlations

* The average age and average risk score of beneficiaries exhibits weak overall correlations with other financial metrics [total charge amount, total allowed amount, total Medicare payment amount and total Medicare standard amount] which indicates that age and risk may not significantly impact payment amounts.

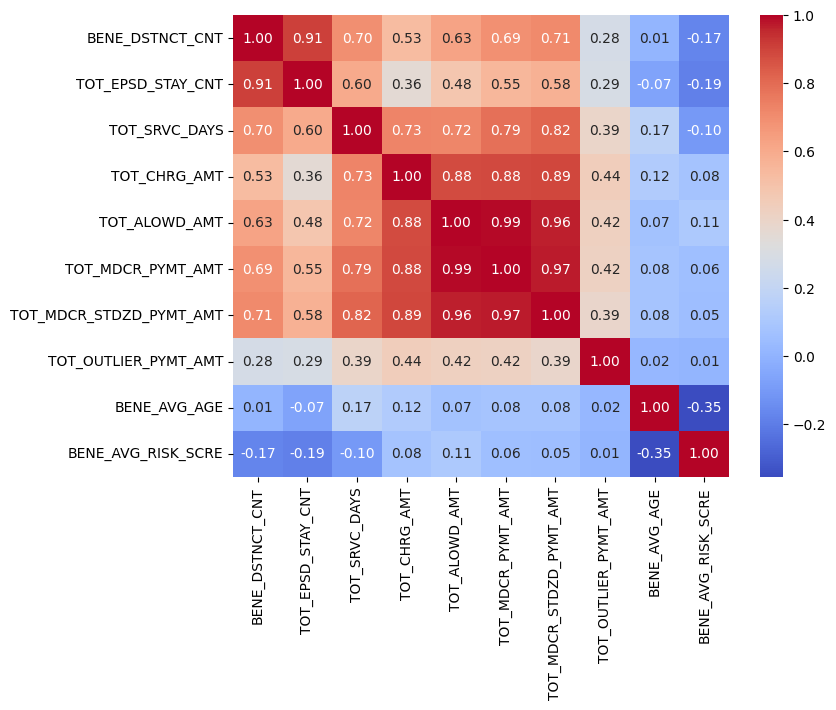


Figure 14: Correlation Heatmap for Providers’ Data

A correlation matrix has been generated to analyze the correlation between the Medicare payment amount and service minutes administered at various facilities, as shown in Figure 15. And it has been found that there is a strong correlation between the payment amount and service minutes.

A red squares with black text

AI-generated content may be incorrect.

Figure 15: Correlation Heatmap for Medicare Payment and Visits

The relationship between the total Medicare payment amount (TOT\_MDCR\_PYMT\_AMT) and count of visits; nursing visits (NRSING\_VISITS\_CNT), social work visits (MSW\_VISITS\_CNT), and home health aide visits (AIDE\_VISITS\_CNT) are analyzed since those frequencies of visits may potentially influence the payment amount. Figure 16 shows the relationship with nursing visits, while Figure 17 shows the same plot but excludes outliers to enlarge the scatter plot concentrated on the left side in detail. Figures 18 and 20 show the correlation with social work and home health aide visits respectively, and Figures 19 and 21 remove outliers. Overall, all the plots show a trend of positive correlation between the payment amounts and each number of visits. It suggests that all three variables related to visits may be used as predictors to forecast the total Medicare payment amount.

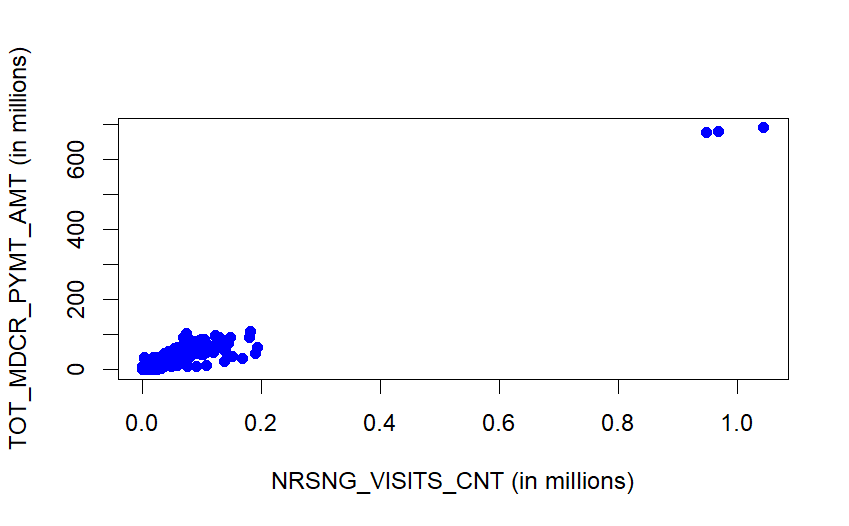


Figure 16: Nursing Visits (incl. Outliers)

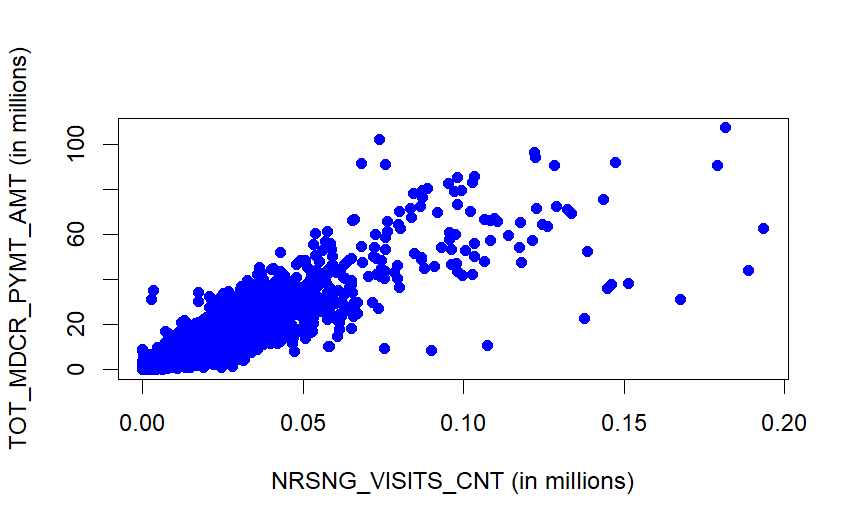


Figure 17: Nursing Visits (excl. Outliers)

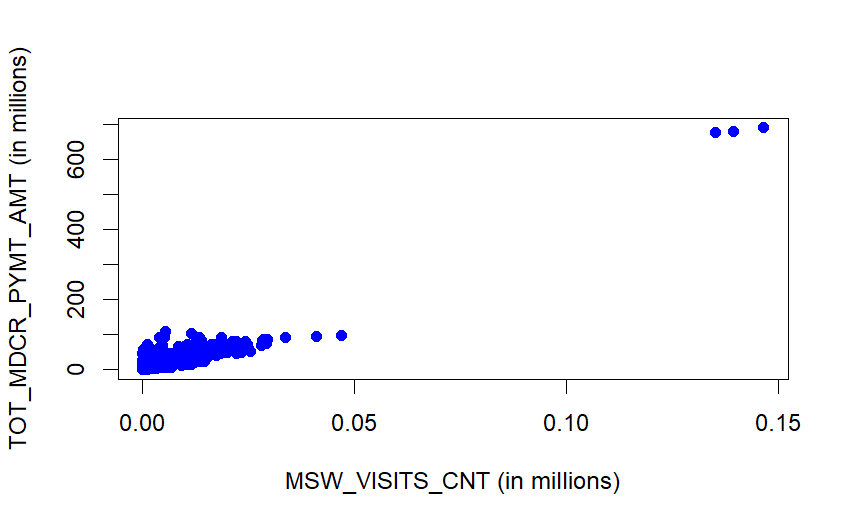


Figure 18: Social Work Visits (incl. Outliers)

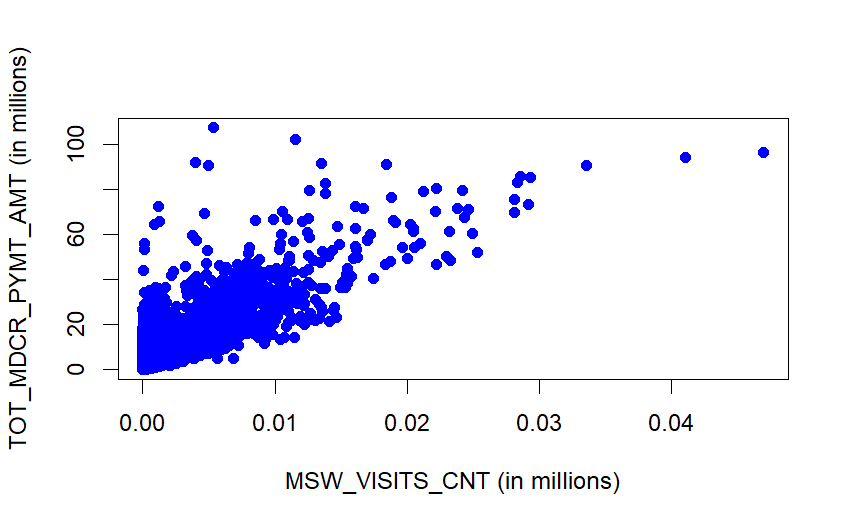


Figure 19: Social Work Visits (excl. Outliers)

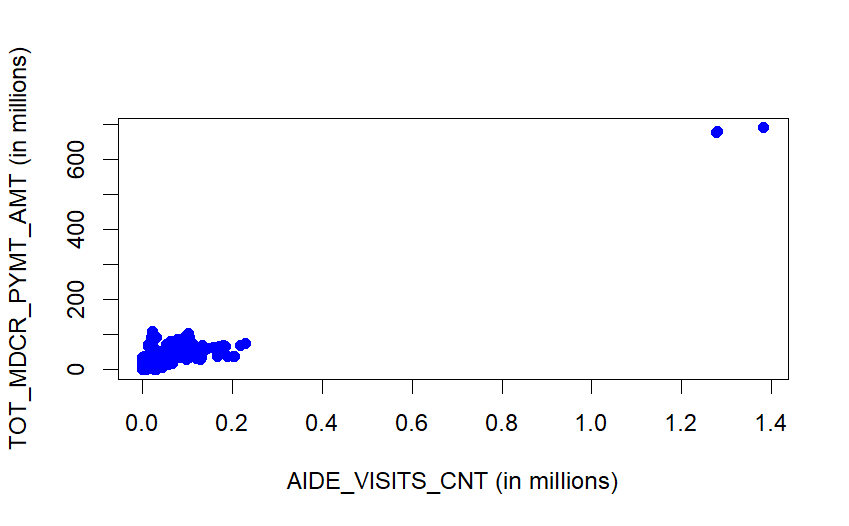


Figure 20: Aide Visits (incl. Outliers)

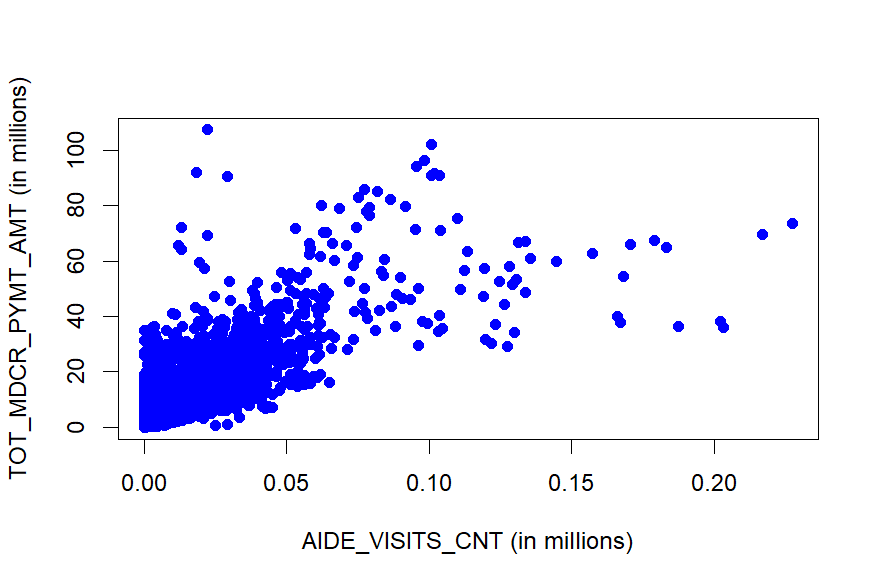


Figure 21: Aide Visits (excl. Outliers)

#### 

#### **4.3.1 Dealing with missing values**

All the variables being considered for analysis were of character data type, most of the columns had special characters (\*). The following steps were followed to resolve the data issues and generate the boxplots, correlations, graphs, heatmaps, and PCA (Principal Component Analysis) analysis was done to reduce the dimensionality to analyze the data and gain an understanding of how the data can be used for further analysis.

* Rows were checked for special characters (\*), and when found were replaced with NA.
* The data type was converted to numeric for conducting further analysis, i.e. running exhaustive search, predictive models.
* For performing PCA (Principal Component Analysis) on chronic condition variables, we lack Hospice service category data, where we replaced N/A to 0 to make data more meaningful, and suitable for performing PCA.

#### **4.3.2 Data exclusion**

The data set being analyzed includes several variables that do not present any relevant information for the analysis or have no information. Table 5 lists out the columns that are being dropped from being considered for our analysis as the data in these columns is either not directly relevant to our analysis or most of the rows have N/A values as per the EDA that was performed.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Exclusion reason** |
| YEAR\_TYPE | Identifies if the data is for calendar year (CY) or fiscal year (FY). | Information is not relevant to the analysis. |
| PRVDR\_ID | The 6-digit CCN for the provider. |
| PRVDR\_CITY | The city where the provider is located |
| PRVDR\_ZIP | The provider’s ZIP code |
| BENE\_MALE\_PCT | Percent of beneficiaries who are male. | Gender and racial background information is not relevant to the analysis |
| BENE\_FEML\_PCT | Percent of beneficiaries who are female. |
| BENE\_RACE\_WHT\_PCT | Percent of beneficiaries who are non-Hispanic white. |
| BENE\_RACE\_BLACK\_PCT | Percent of beneficiaries who are a non-Hispanic black or African American. |
| BENE\_RACE\_API\_PCT | The percentage of beneficiaries are Asian Pacific Islander. |
| BENE\_RACE\_HSPNC\_PCT | Percent of beneficiaries who are Hispanic. |
| BENE\_RACE\_NATIND\_PCT | Percent of beneficiaries who are American Indian or Alaska Native. |
| BENE\_RACE\_UNK\_PCT | Percent of beneficiaries with an unknown race. |
| BENE\_RACE\_OTHR\_PCT | Percent of beneficiaries which are of another race. |
| INDVDL\_PT\_MNTS | Number of physical therapy minutes administered to the resident individually by the SNF/IRF provider | Most of the rows have N/A values. |
| CNCRNT\_GRP\_PT\_MNTS | Number of physical therapy minutes administered to the resident concurrently with one other resident or as part of a group of residents by the SNF/IRF provider |
| COTRT\_PT\_MNTS | Number of physical therapy minutes administered to the resident concurrently with another discipline by the IRF provider |
| INDVDL\_OT\_MNTS | Number of occupational therapy minutes administered to the residents individually by the SNF/IRF provider |
| CNCRNT\_GRP\_OT\_MNTS | Number of occupational therapy minutes administered to the resident concurrently with one other resident or as part of a group of residents by the SNF/IRF provider |
| COTRT\_OT\_MNTS | Number of occupational therapy minutes administered to the resident concurrently with another discipline by the IRF provider. |
| **Column Name** | **Description** |
| INDVDL\_SLP\_MNTS | Number of speech-language pathology therapy minutes administered to the residents individually by the SNF/IRF provider. |
| CNCRNT\_GRP\_SLP\_MNTS | Number of speech-language pathology therapy minutes administered to the resident concurrently with one other resident or as part of a group of residents by the SNF/IRF provider. |
| COTRT\_SLP\_MNTS | Number of speech-language pathology therapy minutes administered to the residents concurrently with another discipline by the IRF provider. |

Table 5: Variable Exclusion

#### **4.3.3 Outlier Analysis**

Analysis for outliers has been conducted on key data elements that will be leveraged for further analysis to understand the range in which most of the records fall for the provider’s level data for all three years. Variables analyzed for provider level information are beneficiary count (Figure 22), total episode or stay count (Figure 23) and total service days (Figure 24). Our analysis did not find any outliers that were too far from the interquartile range, so outliers will not be excluded for further analysis.

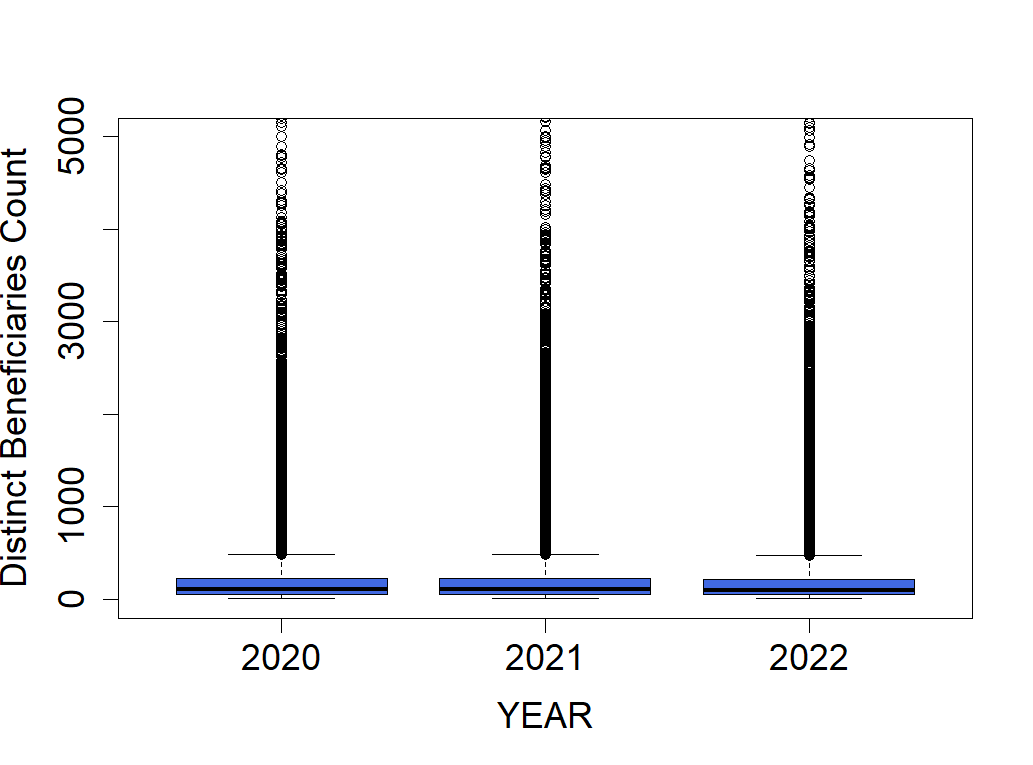


Figure 22: Boxplot for Distinct Beneficiary Count

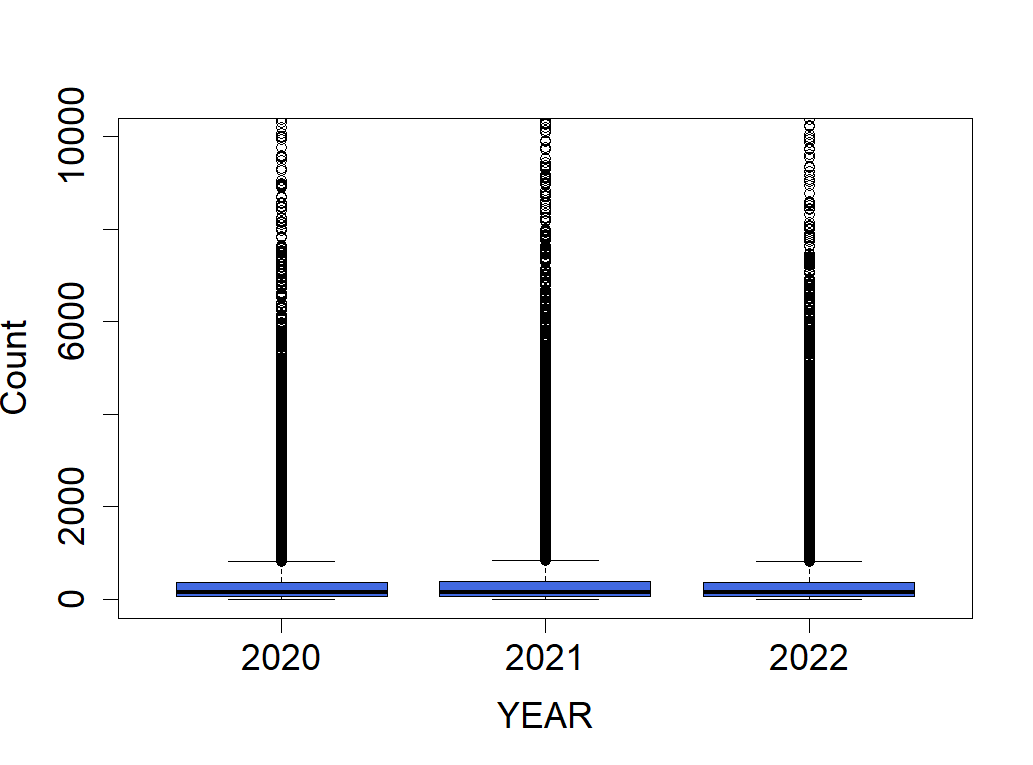


Figure 23: Episode or Stay Count

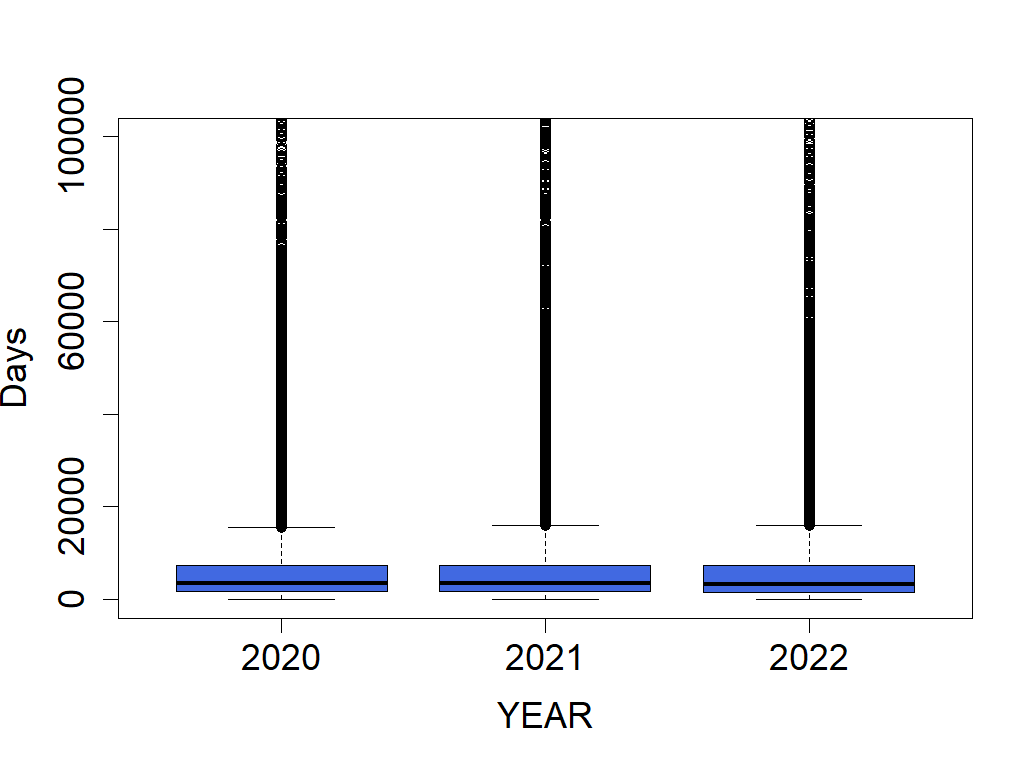


Figure 24: Days of Service

The variables analyzed for outliers at the national level data across the 3 years are, amount charged by the providers (Figure 25), the total amount allowed by Medicare (Figure 26), the total standardized payment amount (Figure 27) and the total Medicare payment amount (Figure 28). The interquartile range (IQR) remained close across the three years. And no data points were too far from the IQR. We are considering running the models with all the data at this point.

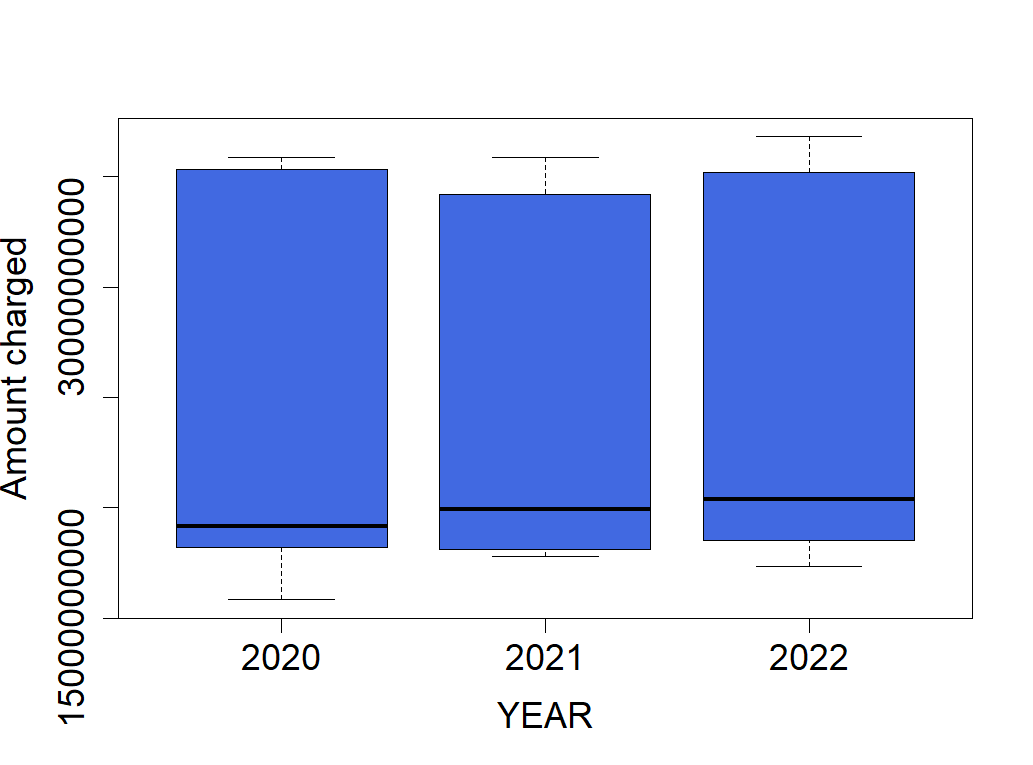


Figure 25: Total Charge Amount

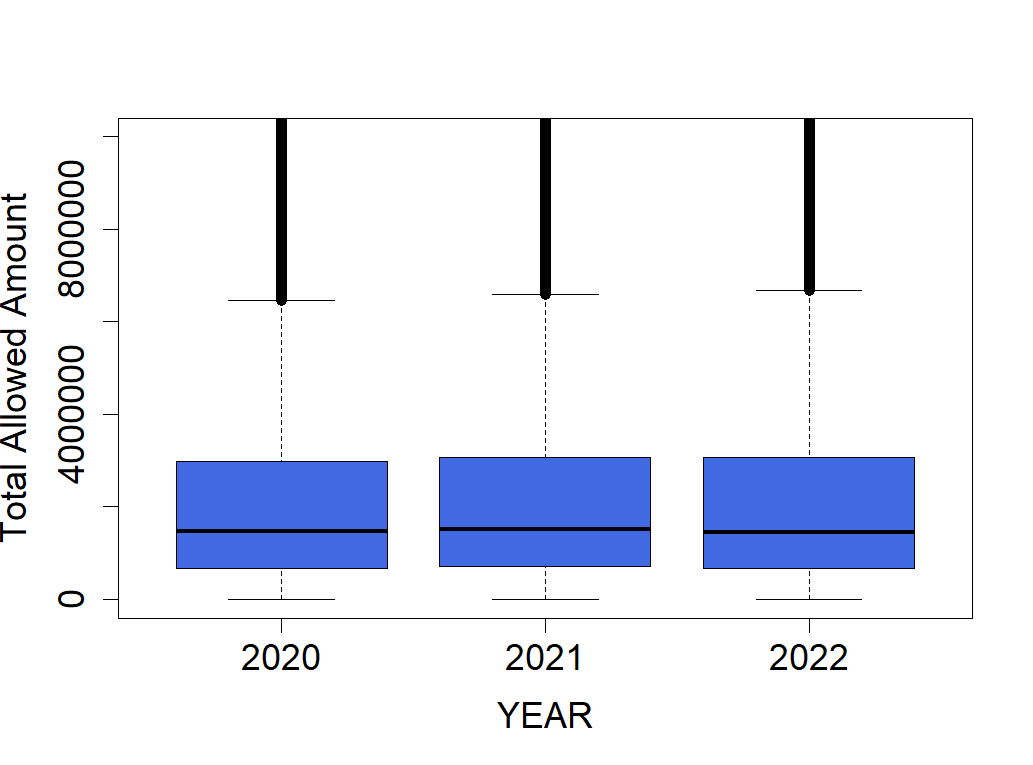


Figure 26: Total Amount Allowed by Medicare

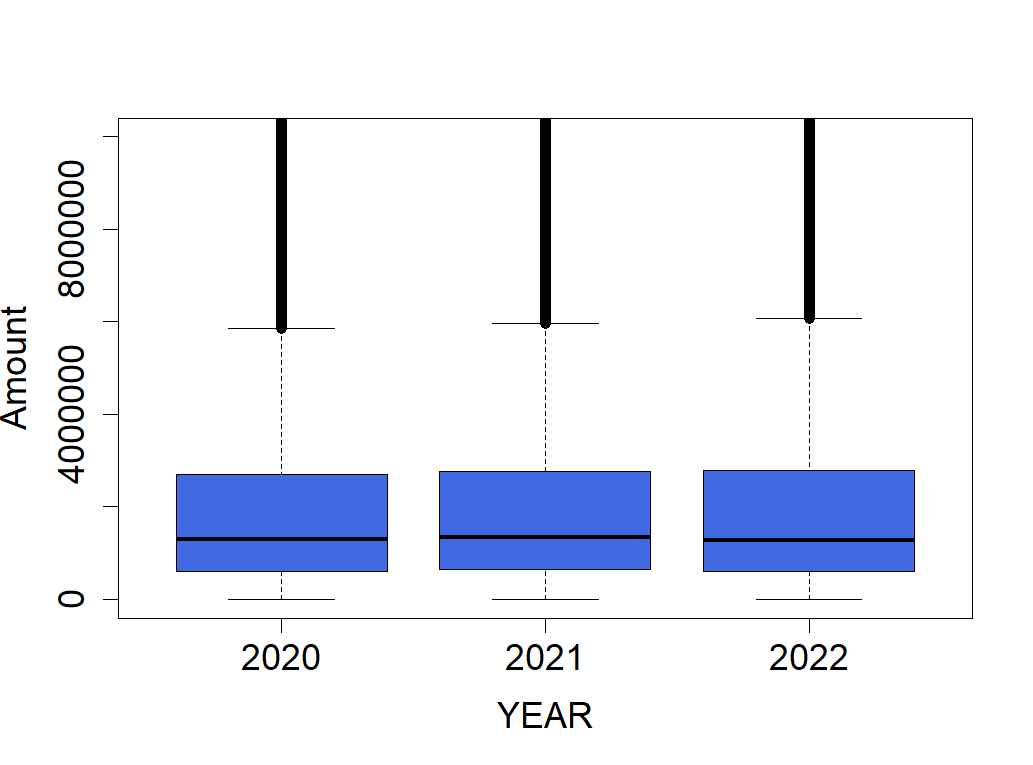


Figure 27: Total Medicare Payment Amount

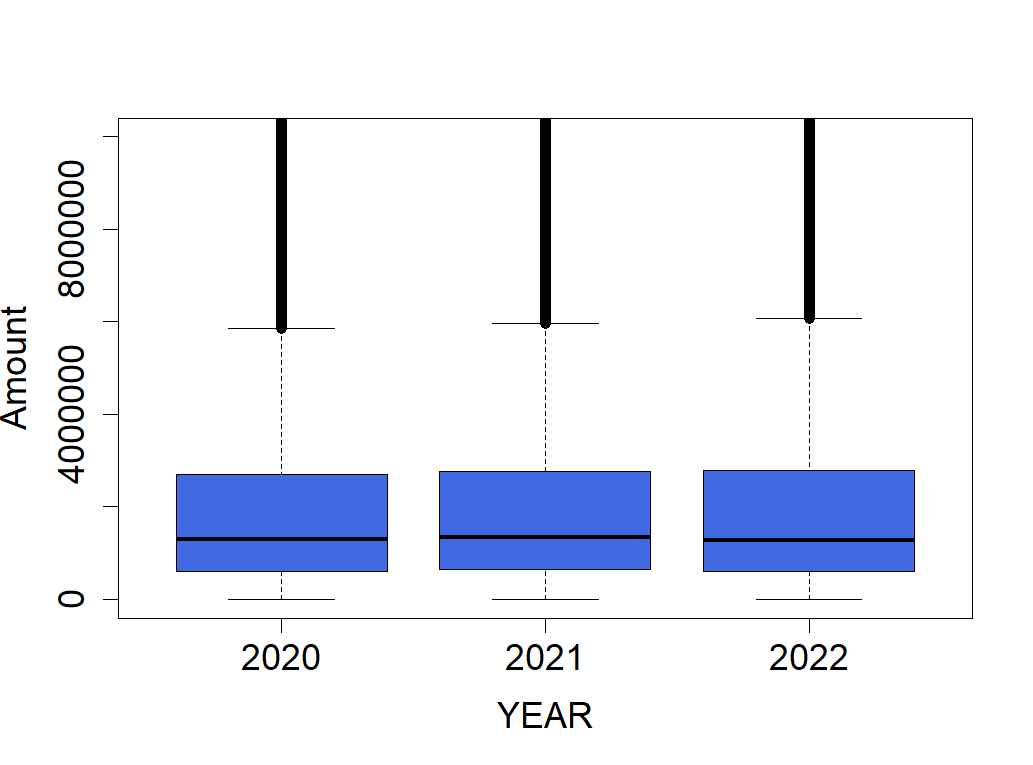


Figure 28: Total Medicare Standard Payment

#### **4.3.4 Dimension Reduction**

According to our initial understanding, two types of variables 1) Percent of beneficiaries identified with chronic conditions and 2) Percent of episodes with a primary diagnosis of disease can be potential predictors of total Medicare payment amount. However, in the dataset, there are 25 variables related to chronic conditions and 15 variables related to diagnosis episodes. Since there are too many variables, we conduct dimension reduction by using Principal Component Analysis (PCA). PCA analysis is conducted for each type of column, chronic conditions and primary diagnosis episodes respectively. After the implementation, we identified principal components that contain most of the variances of the original variables.

##### **PCA Analysis for Percent of beneficiaries identified with chronic conditions**

According to the results listed in Table 6, PC1 is the most important component, explaining 63.57% of the variance. By PC4, the cumulative proportion reaches 79.35%, which is almost 80%. This implies that the first 4 components capture most of the meaningful variance, suggesting that dimensionality reduction may be effective using only these components. Next, Figure 29 shows the scree plot from PC1 through PC25. It visualizes the proportion of the variables and it’s obvious that the first component is significant compared to the remaining components. Finally, Figure 30 illustrates how much each variable contributes to PC1 by plotting the top 10 variables that influence the most strongly to the component. It shows that each chronic condition contributes almost equivalently to PC1.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 |
| Standard Deviation | 3.9865 | 1.3667 | 1.0954 | 0.9369 | 0.8131 | 0.7987 | 0.7153 | 0.6480 | 0.5944 | 0.5709 |
| Proportion of variance | 0.6357 | 0.0747 | 0.0480 | 0.0351 | 0.0264 | 0.0255 | 0.0205 | 0.0168 | 0.0141 | 0.0130 |
| Cumulative proportion | 0.6357 | 0.7104 | 0.7584 | 0.7935 | 0.8200 | 0.8455 | 0.8660 | 0.8828 | 0.8969 | 0.9099 |
|  | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 |
| Standard Deviation | 0.5563 | 0.5477 | 0.5032 | 0.4773 | 0.4468 | 0.4014 | 0.3914 | 0.3508 | 0.3348 | 0.3211 |
| Proportion of variance | 0.0124 | 0.0120 | 0.0101 | 0.0091 | 0.0080 | 0.0065 | 0.0061 | 0.0049 | 0.0045 | 0.0041 |
| Cumulative proportion | 0.9223 | 0.9343 | 0.9444 | 0.9535 | 0.9615 | 0.9680 | 0.9741 | 0.9790 | 0.9835 | 0.9876 |
|  | PC21 | PC22 | PC23 | PC24 | PC25 |  |  |  |  |  |
| Standard Deviation | 0.2891 | 0.2776 | 0.2646 | 0.2376 | 0.1488 |  |  |  |  |  |
| Proportion of variance | 0.0033 | 0.0031 | 0.0028 | 0.0023 | 0.0009 |  |  |  |  |  |
| Cumulative proportion | 0.9910 | 0.9941 | 0.9969 | 0.9991 | 1.0000 |  |  |  |  |  |

Table 6: Importance of components for Chronic Condition Variables

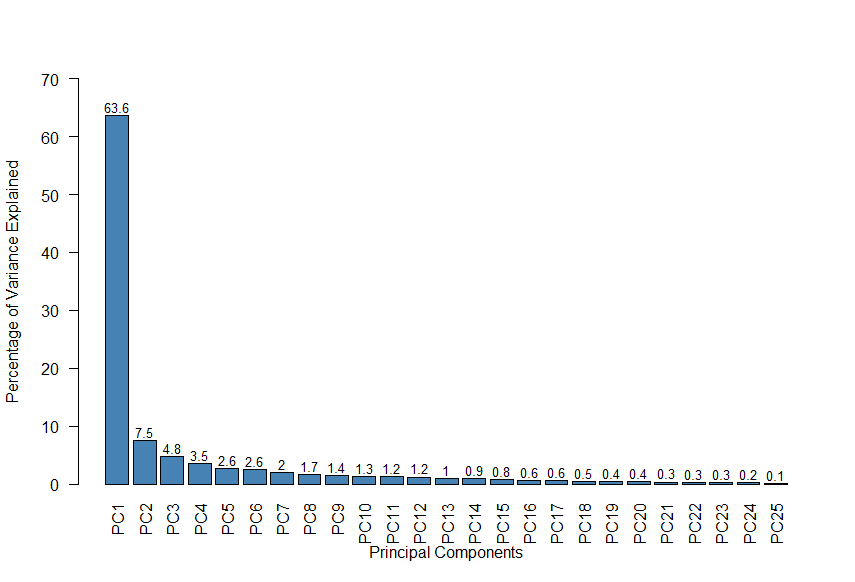


Figure 29: Scree Plot of PCA for Percent of Beneficiaries Identified with Chronic Condition

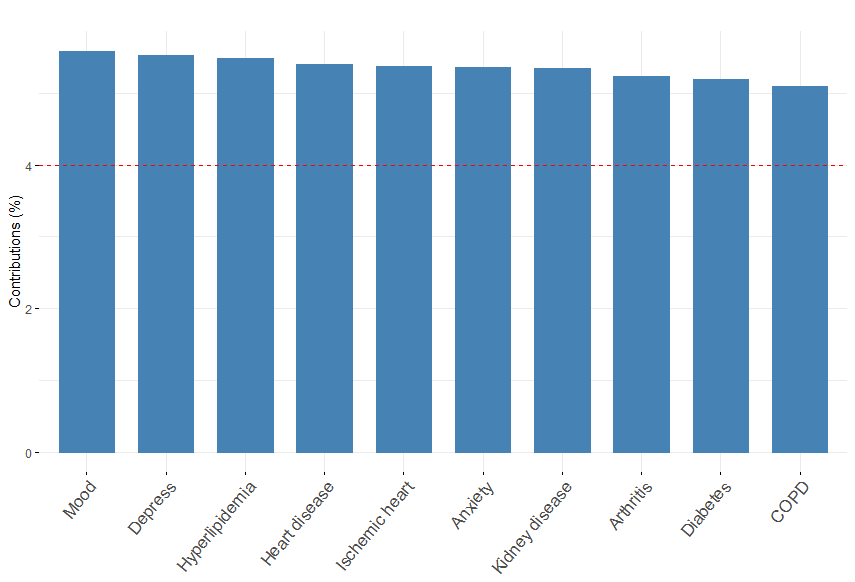


Figure 30: Contribution of Variables to PC1 Chronic Conditions

##### **PCA Analysis for Percent of episodes with a primary diagnosis of disease**

According to the results listed in Table 7, PC1 is the most important component, explaining 14.58% of the variance. PC2 and PC3 are also significant as they are 11.43% and 8.92% respectively. The cumulative proportion shows that the first 3 PCs account for 34.92 % of the variance. By PC10, the cumulative proportion reaches 79.07 % which captures 80% of the variance. This implies that the first 10 components capture most of the meaningful variance, suggesting that dimensionality reduction may be effective using only these components. Next, Figure 31 shows the scree plot from PC1 through PC15. It visualizes the proportion of the variables and it’s obvious that the first three components are significant compared to the remaining components. Finally, Figure 32 illustrates how much each variable contributes to PC1 by plotting 15 variables. It tells us that “Diseases of the Circulatory System” and “Injury, Poisoning and certain other consequences of external causes & external causes of Morbidity” are most influential in explaining the PC1’s structure. On the other hand, it shows that the least 4 variables that are “Endocrine”, “Respiratory System”, “Diseases of the Eye and Adnexa & Disease of the Ear and Mastoid process” and “Pregnancy, Childbirth and the Puerperium” play less of roll in PC1.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 |
| Standard Deviation | 1.4787 | 1.3091 | 1.1566 | 1.0362 | 1.0045 | 0.9993 | 0.9804 | 0.9599 | 0.9350 | 0.8853 |
| Proportion of variance | 0.1458 | 0.1143 | 0.0892 | 0.0716 | 0.0673 | 0.0666 | 0.0641 | 0.0614 | 0.0583 | 0.0523 |
| Cumulative proportion | 0.1458 | 0.2600 | 0.3492 | 0.4208 | 0.4881 | 0.5546 | 0.6187 | 0.6801 | 0.7384 | 0.7907 |
|  | PC11 | PC12 | PC13 | PC14 | PC15 |  |  |  |  |  |
| Standard Deviation | 0.8736 | 0.8583 | 0.8059 | 0.7658 | 0.6359 |  |  |  |  |  |
| Proportion of variance | 0.0509 | 0.0491 | 0.0433 | 0.0391 | 0.0270 |  |  |  |  |  |
| Cumulative proportion | 0.8415 | 0.8907 | 0.9340 | 0.9730 | 1.0000 |  |  |  |  |  |

Table 7: Importance of Components for Primary Diagnosis Variables

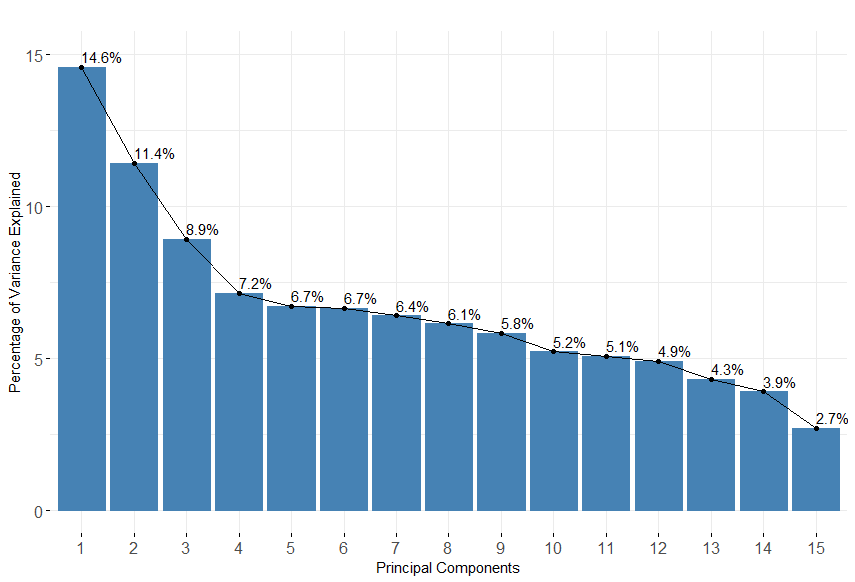


Figure 31: Scree Plot of PCA for Percent of Episodes with a Primary Diagnosis of Disease

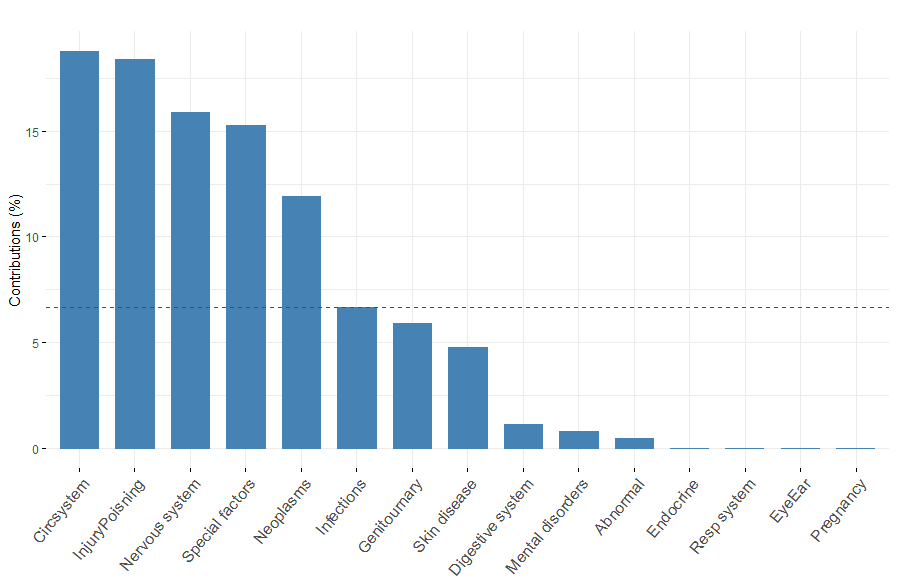


Figure 32: Contribution of Variables to PC1 for Primary Diagnosis

## **5. Methodologies**

As per the data analysis conducted so far, we outline the methodologies included in our research to analyze the total Medicare payment amount for various post-acute care (PAC) facilities. The methodologies include Principal Component Analysis (PCA) for dimensionality reduction, an exhaustive search for variable selection, and the application of multiple linear regression and elastic net regression models.

### **5.1 Principal Component Analysis (PCA)**

**PCA is a statistical technique used for dimensionality reduction that transforms a dataset into a set of orthogonal (uncorrelated) variables called as principal components which retain as much variance as possible. This technique helps to identify the most significant variables that contribute to total Medicare payment amounts. PCA is particularly useful in diminishing complex datasets without compromising essential information, making it a valuable tool in our analysis of Medicare costs as mentioned in,** [26]**.**

### **5.2 Exhaustive search for variable selection**

**An exhaustive search was conducted to identify the optimal subset of variables for the regression models by evaluating all possible combinations. The selection criteria focused on maximizing the adjusted R-squared value and minimizing the Akaike Information Criterion (AIC), as mentioned in,** [27]**.**

### **5.3 Multiple linear regression**

**A regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line. There are several types of multiple regression analyses (e.g. standard, hierarchical, set wise, stepwise) only two of which will be presented here (standard and stepwise). Which type of analysis is conducted depends on the question of interest to the researcher. The model's performance is evaluated using metrics like R-squared, RMSE, and MAE to assess how well the independent variables explain the variation in the dependent variable. Multiple regression can also suffer from overfitting, which is when your model fits the data too well and loses its ability to generalize to new or unseen data. As per​** [28]**, overfitting can occur when you have too many independent variables, or when your variables are highly correlated with each other.**

### **5. 4 Ridge / Lasso / Elastic Net Regression**

**These are techniques that are used to prevent overfitting and perform variable selection when dealing with many predictors. Elastic Net is particularly useful when there are many features as per​** [29]**, some of which might be correlated, as it combines the penalties of both Lasso (L1) and Ridge (L2) regression.**

## **6. Results**

Two different methods for predictive analytics were used, multiple linear regression and elastic net regression. When compared to the results for both the methodologies, multiple liner regression resulted in better performance. There are two patterns of analysis for multiple linear regression, one implemented one model using all records in the dataset, and the other generated five models using the dataset of each service category (SNF, LTC, IRF, HH, and Hospice), as the total Medicare payment amount is related to the services offered by the facilities.  

**6.1 Multiple Linear Regression (for all service categories)**

**Before running the model, 16 variables were selected by eliminating irrelevant variables and cost-related variables that have multicollinearity problems. In addition to that, four PC components and ten PC components were selected from chronic condition and primary diagnosis episodes, respectively. In total 30 variables were selected to run exhaustive search. As a result of the exhaustive search, 8 variables in Table 8 were selected as this combination of the model has the highest adjusted R square and the lowest AIC value.**

**After running the multiple linear regression, coefficients in Table 8 were identified. The adjusted R square of the model is 0.9334. It suggests that 93.34% of the variation in the target variable is explained by these predictors, which is considered a good model fit. Table 9 shows the key performance metrics for evaluating the model’s prediction. According to this, a high MAPE value (60.73%) indicates that the model may have high errors when predicting,** this explains the high RMSE that we see in Table 9. 

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | **311,374.74** |
| **BENE\_DSTNCT\_CNT** | **5,515.37** |
| **TOT\_SRVC\_DAYS** | **206.94** |
| **NRSNG\_VISITS\_CNT** | **-161.46** |
| **MSW\_VISITS\_CNT** | **-896.89** |
| **TOT\_PT\_MNTS** | **-7.63** |
| **TOT\_OT\_MNTS** | **26.36** |
| **TOT\_HH\_LUPA\_EPSDS\_CNT** | **-12,779.29** |
| **PC1\_B** | **-121,311.23** |

Table 8: Coefficients of Multiple Linear Regression Model for all service categories

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | 3,826 | Model overpredicts the target by about 3.8K. |
| RMSE | 1,615,535 | Predictions are off by around 1.62 million on average.  (The high RMSE might be due to outliers) |
| MAE | 730,076 | On average, predictions are off by about 730K from the actual Medicare payment amount. |
| MPE | -22.26% | Model underpredicts by about 22% on average. |
| MAPE | 60.73% | On average, predictions are off by about 61% |

Table 9: Prediction Performance of Multiple Linear Regression Model for all service categories

### **6.2 Multiple Linear Regression (by each service category)**

Before conducting linear regression for each service category, outliers were addressed using a capping methodology. Subsequently, a Principal Component Analysis (PCA) was performed on the selected variables, followed by an exhaustive search to identify the most relevant variables. This process resulted in the selection of eight key variables for the multiple linear regression analysis. It is important to note that the specific variables may vary for each service category, as detailed in the subsequent steps.

#### **6.2.1 HH Service Category**

**After running the multiple linear regression for Home Health service category, coefficients in Table 10 were identified. The adjusted R square of the model is 0.9620. It suggests that 96.2% of the variation in the target variable is explained by these predictors, which is considered a good model fit. Table 11 shows the key performance metrics for evaluating the model’s prediction. According to this, an average MAPE value (26.9%) indicates that the model may have moderate errors while predicting.**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | -235,500.00 |
| **BENE\_DSTNCT\_CNT** | 2,335.00 |
| **TOT\_SRVC\_DAYS** | 215.40 |
| **BENE\_DUAL\_PCT** | 4,586.00 |
| NRSNG\_VISITS\_CNT | -136.60 |
| MSW\_VISITS\_CNT | 574.80 |
| TOT\_NRSNG\_MNTS | 1.87 |
| TOT\_PT\_MNTS | -2.04 |
| TOT\_OT\_MNTS | -3.26 |

Table 10: Coefficients of Multiple Linear Regression Model for HH

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | 9,222 | Model overpredicts the target by about 9.22K. |
| RMSE | 525,463 | Predictions are off by around 525K on average. |
| MAE | 254,789 | On average, predictions are off by about 255K from the actual Medicare payment amount. |
| MPE | -1.96% | Model underpredicts by about 1.96% on average. |
| MAPE | 26.9% | On average, predictions are off by about 26.9%. |

Table 11: Prediction Performance of Multiple Linear Regression Model for HH Records

#### **6.2.2 HOS Service Category**

**After running the multiple linear regression for Hospice service category, coefficients in Table 12 were identified. The adjusted R square of the model is 0.9741. It suggests that 97.41% of the variation in the target variable is explained by these predictors, which is considered a good model fit. Table 13 shows the key performance metrics for evaluating the model’s prediction. According to this, an average MAPE value (19.18%) indicates that the model may have moderate errors when predicting.**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | -568,100.00 |
| **BENE\_DSTNCT\_CNT** | 3,905.00 |
| **TOT\_EPSD\_STAY\_CNT** | 450.40 |
| **TOT\_SRVC\_DAYS** | 125.90 |
| MSW\_VISITS\_CNT | -152.40 |
| TOT\_NRSNG\_MNTS | 0.93 |
| PC1\_P | 166,700.00 |
| PC3\_P | -168,000.00 |
| PC10\_P | 132,300.00 |

Table 12: Coefficients of Multiple Linear Regression Model for HOS

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | -1,083 | Model underpredicts the target by about 1.08K. |
| RMSE | 764,216 | Predictions are off by around 764K on average. |
| MAE | 426,632 | On average, predictions are off by about 426K from the actual Medicare payment amount. |
| MPE | 5.81% | Model overpredicts by about 5.81% on average |
| MAPE | 19.18% | On average, predictions are off by about 19.18%. |

Table 13: Prediction Performance of Multiple Linear Regression Model for HOS Records

#### **6.2.3 SNF Service Category**

**After running the multiple linear regression for Skilled Nursing Facilities service category, coefficients in Table 14 were identified. The adjusted R square of the model is 0.9282. It suggests that 92.82% of the variation in the target variable is explained by these predictors, which is considered a good model fit. Table 15 shows the key performance metrics for evaluating the model’s prediction. According to this, an average MAPE value (18.84%) indicates that the model may have moderate errors when predicting.**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | -285,800.00 |
| **TOT\_SRVC\_DAYS** | 569.00 |
| **BENE\_DUAL\_PCT** | 2,465.00 |
| **BENE\_AVG\_RISK\_SCRE** | 65,550.00 |
| PC1\_B | 72,020.00 |
| PC3\_B | -69,720.00 |
| PC5\_P | -45,240.00 |
| PC7\_P | 33,400.00 |
| PC10\_P | -51,930.00 |

Table 14: Coefficients of Multiple Linear Regression Model for SNF Records

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | -2,629 | Model underpredicts the target by about 2.63K |
| RMSE | 508,634 | Predictions are off by around 509K on average. |
| MAE | 289,377 | On average, predictions are off by about 289K from the actual Medicare payment amount. |
| MPE | 1.40% | Model overpredicts by about 1.40% on average. |
| MAPE | 18.84% | On average, predictions are off by about 18.84%. |

Table 15: Prediction Performance of Multiple Linear Regression Model for SNF Records

#### **6.2.4 LTC Service Category**

**After running the multiple linear regression, coefficients in table 16 were identified. The adjusted R square of the model is 0.8813. It suggests that 88.13% of the variation in the target variable is explained by these predictors, which is considered good model fit. Table 17 shows the key performance metrics for evaluating the model’s prediction. According to this, an average MAPE value (22.46%) indicates that the model may have moderate errors when predicting.**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | -169,376.70 |
| **BENE\_DSTNCT\_CNT** | 41,814.07 |
| **TOT\_EPSD\_STAY\_CNT** | -36,664.49 |
| **TOT\_SRVC\_DAYS** | 1,168.87 |
| **BENE\_DUAL\_PCT** | 31,902.31 |
| PC2\_B | 387,133.04 |
| PC3\_B | -284,254.02 |
| PC4\_B | 334,209.19 |
| PC10\_P | -668,464.38 |

Table 16: Coefficients of Multiple Linear Regression Model for LTC Records

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | 186,785 | Model overpredicts the target by about 186.79K. |
| RMSE | 1,768,454 | Predictions are off by around 1.77M on average. |
| MAE | 1,296,739 | On average, predictions are off by about 1.30M from the actual Medicare payment amount. |
| MPE | -10.49% | Model underpredicts by about 10.49% on average. |
| MAPE | 22.46% | On average, predictions are off by about 22.46%. |

Table 17: Prediction Performance of Multiple Linear Regression Model for LTC Records

#### **6.2.5 IRF Service Category**

**After running the multiple linear regression for Inpatient Rehabilitation facilities, coefficients in Table 18 were identified. The adjusted R square of the model is 0.9184. It suggests that 91.84% of the variation in the target variable is explained by these predictors, which is considered a good model fit. Table 19 shows the key performance metrics for evaluating the model’s prediction. According to this, an average MAPE value (21.67%) indicates that the model may have moderate errors when predicting.**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | 411,400.00 |
| **BENE\_DSTNCT\_CNT** | 18,080.00 |
| **TOT\_**EPSD\_STAY\_CNT | -13,960.00 |
| TOT\_SRVC\_DAYS | 404.40 |
| TOT\_PT\_MNTS | 10.04 |
| TOT\_OT\_MNTS | 6.60 |
| TOT\_SLP\_MNTS | 18.29 |
| PC2\_B | 259,200.00 |
| PC3\_B | -454,200.00 |

Table 18: Coefficients of Multiple Linear Regression Model for IRF Records

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | 63,894 | Model overpredicts the target by about 63.89K. |
| RMSE | 1,751,417 | Predictions are off by around 1.75M on average. |
| MAE | 1,144,516 | On average, predictions are off by about 1.14M from the actual Medicare payment amount. |
| MPE | -9.24% | Model underpredicts by about 9.24% on average. |
| MAPE | 21.67% | On average, predictions are off by about 21.67%. |

Table 19: Prediction Performance of Multiple Linear Regression Model for IRF Records

### **6.3 Elastic Net Regression (for all records)**

As mentioned in 6.1, the same 8 variables were used for elastic net regression. Table 20 shows the coefficients of each variable, and Table 21 shows the model’s prediction performance. **The adjusted R square of the model is 0.9197. It suggests that 91.97% of the variation in the target variable is explained by these predictors, which is considered a good model fit. According to Table 21, the model seems to have relatively high errors as mentioned by high MPE (-23.50%) and MAPE (61.13%).**

|  |  |
| --- | --- |
| **Variables** | **Coefficients** |
| **(Intercept)** | 322,147.00 |
| **BENE\_DSTNCT\_CNT** | 5,514.51 |
| **TOT\_SRVC\_DAYS** | 200.47 |
| **NRSNG\_VISITS\_CNT** | -148.32 |
| **MSW\_VISITS\_CNT** | -832.48 |
| **TOT\_PT\_MNTS** | -7.45 |
| **TOT\_OT\_MNTS** | 26.19 |
| **TOT\_HH\_LUPA\_EPSDS\_CNT** | -13,022.59 |
| **PC1\_B** | -122,263.01 |

Table 20: Coefficients of Elastic Net Regression

|  |  |  |
| --- | --- | --- |
| **Metric** | **Value** | **Description** |
| ME | 3,851 | Model overpredicts the target by about 3.9K. |
| RMSE | 1,615,874 | Predictions are off by around 1.61 million on average.  (The high RMSE might be due to outliers) |
| MAE | 730,020 | On average, predictions are off by about 730K from the actual Medicare payment amount. |
| MPE | -23.50% | Model underpredicts by about 24% on average. |
| MAPE | 61.13% | On average, predictions are off by about 61% |

Table 21: Prediction Performance of Elastic Net Regression

### **6.4 Model comparison**

**6.4.1 Multiple Linear Regression vs Elastic Net Regression (for all service categories)**

     Table 22 shows the performance comparison between multiple linear regression and elastic net regression models. As for the accuracy of the prediction, both models have similar values across all metrics, but overall multiple linear regression model has slightly better performance than elastic net regression. In terms of goodness of fit, multiple linear regression model represents slightly better adjusted R square value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Multiple Linear Regression** | | **Elastic Net Regression** | |
| **Value** | **vs Elastic** | **Value** | **vs Linear** |
| ME | 3,826 | Better | 3,851 | Worse |
| RMSE | 1,615,535 | Better | 1,615,874 | Worse |
| MAE | 730,076 | Slightly worse | 730,020 | Slightly better |
| MPE | -22.26% | Better | -23.50% | Worse |
| MAPE | 60.73% | Better | 61.13% | Worse |
| R²adj | 93.34% | Better | 91.97% | Worse |

Table 22: Performance Comparison between Multiple Linear and Elastic Net Regression

#### **6.4.2 Multiple Linear Regression (for all service categories) vs Multiple Linear Regression (by each service category)**

Table 23 shows the performance comparison between the multiple linear regression model used all service categories and the 5 models used the dataset of each service category. Overall, prediction performance of models of HH, HOS, and SNF are better than the model using all service categories across all metrics except ME of HH. Especially, their RMSE values are much better than all records. LTC and IRF show slightly worse RSME and worse ME values than the one using all service categories, but their MPE and MAPE show improvements. In terms of adjusted R square, all models show above 80-90% which is a good model fit.

     In conclusion, predictive models segmented by each service category have better accuracy performance than the model with all service categories in general, although LTC and IRF have some issues as they make large errors in some metrics.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **All service categories** | **HH** | **HOS** | **SNF** | **LTC** | **IRF** |
| ME | 3,826 | 9,222 | -1,083 | -2,629 | 186,785 | 63,894 |
| RMSE | 1,615,535 | 525,463 | 764,216 | 508,634 | 1,768,454 | 1,751,417 |
| MAE | 730,076 | 254,789 | 426,632 | 289,377 | 1,296,739 | 1,144,516 |
| MPE | -22.26% | -1.96% | 5.81% | 1.40% | -10.49% | -9.24% |
| MAPE | 60.73% | 26.90% | 19.18% | 18.84% | 22.46% | 21.67% |
| R²adj | 93.34% | 96.20% | 97.41% | 92.82% | 88.13% | 91.78% |

Table 23: Performance comparison of all service categories with each service category

## **7. Conclusions**

As per the results, which are based on the Multiple Linear Regression Model considering the provider data for all service categories, the following are the findings.

It has been confirmed that a combination of the number of beneficiaries utilizing the facility, total service days, and occupational therapy minutes have a positive influence on the payment, while nursing visits, social work visits, and physical therapy minutes have a negative impact on the payment. The results have also proven that the greater the number of patients utilizing the PAC facilities and the greater number of days they get the services, the Medicare payment increases.

While occupational therapy minutes have a positive impact on the Medicare payment, the Physical therapy minutes have a negative impact. A few other variables that have a negative impact on the Total Medicare payment amount include the low utilization payment amount episodes, number of nursing visits, and the principal component that was an outcome of PCA and exhaustive search analysis. While low utilization payment amount episodes having a negative impact is reasonable, the reason for other attributes having a negative impact on Medicare payment amounts remains unclear.

Further information, and much deeper analysis of these contributing factors, might uncover the reasons contributing to the negative impact.

This can be considered as a shortcoming of this analysis due to lack of data. Additionally, the number of patients and the number of service days having a positive contribution factor toward the payment amount is obvious.

The current data set has many limitations. To highlight patient cost, more information on the services offered by each of the providers, ratings of the providers and data for all the attributes available for all service categories would have been more helpful.

# **Bibliography**

|  |  |
| --- | --- |
| [1] | G. W. Mariam Faizullabhoy, "U.S. Post-acute Care Market – By Service, By Age Group, By Condition, By End Use - Forecast, 2025 to 2034," December, 2024. |
| [2] | C. C. A. M. P. S. Z. L. a. M. S. Jessica P. Lendon, "Overview of Post-acute and Long-term Care Providers and Services Users in the United States," August, 2024. |
| [3] | H. K. White, "Post-acute Care: Current State and Future Directions," pp. 392-395, April, 2019. |
| [4] | B. E. F. J. N. M. N. I. D. Z. D. M. D. P. A. S. H. R. J. John P. Hirdes, "Home Care Quality Indicators (HCQIs) Based on the MDS-HC," 2004. |
| [5] | M. K. M. L. a. M. M. Linda A. Jacobsen, "Overview of Post-acute and Long-term Care Providers and Services Users in the United States," 2011. |
| [6] | D. Dingell, *The Wide Range Of Challenges Facing Senios,* 2011. |
| [7] | K. L. B. M. M. &. K. A. M. Maricruz Rivera-Hernandez, "Seniors don’t use Medicare.Gov: how do eligible beneficiaries obtain information about Medicare Advantage Plans in the United States?," 2021. |
| [8] | V. R. Futchs, "Health Care For The Elderly: How Much? Who Will Pay For It?," 1999. |
| [9] | A. D. H. M. C. R. W. S. Karen M Detering. |
| [10] | S. E. P. Nasiya N. Ahmed, "Acute Care for the Elderly," 2010. |
| [11] | M. S. Joseph G. Ouslander, "The Conundrum of Choosing Post-Acute Care: A Challenge for Patients, Families, and Clinicians," 2019. |
| [12] | A. D. I.-P. J. I. ,. M. Poonam K. Pardasaney, "Measuring Inpatient Rehabilitation Facility Quality of Care: Discharge Self-Care Functional Status Quality Measure," 2018. |
| [13] | J. C. F. Nicholas G. Castle, "What Is Nursing Home Quality and How Is It Measured," 2010. |
| [14] | S. S. Andrew Olenski, "Estimating Nursing Home Quality with Selection," 2024. |
| [15] | J. K. S. R. A. Anthony Williams, "The Nursing Home Five Star Rating: How Does It Compare to Resident and Family Views of Care?," 2014. |
| [16] | M. R. K. S. T. R. T. K. V. M. Margot L Schwartz, "Consumer selection and home health agency quality and patient experience stars," 2021. |
| [17] | E. C. N. R. T. K. D. P. Rachel M. Werner, "Do consumers respond to publicly reported quality information? Evidence from nursing homes," 2012. |
| [18] | E. A. W. D. H. S.-j. M. C. L. A. V. P. E. A. C. R. S. A. A. G. Robert E. Burke, "Hospital Readmission From Post-Acute Care Facilities: Risk Factors, Timing, and Outcomes," 2016. |
| [19] | N. Z. Suzanne R. O'Brien, "Association Between Therapy Intensity and Discharge Outcomes in Aged Medicare Skilled Nursing Facilities Admissions," 2018. |
| [20] | T.-H. L. K.-P. C. Y.-C. T. Ya-Hui Hsu, "Determining the factors influencing the selection of post-acute care models by patients and their families: a qualitative content analysis," 2023. |
| [21] | D. A. T. ,. V. M. Emily A. Gadbois, "Selecting a Skilled Nursing Facility for Postacute Care: Individual and Family Perspectives," 2017. |
| [22] | J. J. E. L. A. L. R. A. C. L. B. L. D. D. M. R. A. E. C. Robert E. Burke, "Evaluating the Quality of Patient Decision-Making Regarding Post-Acute Care," 2018. |
| [23] | H. L. K. M. N. Judith Feder, "Long-Term Care In The United States: An Overview," 2000. |
| [24] | D. R. Zimmerman, "Improving nursing home quality of care through outcomes data: the MDS quality indicators," 2003. |
| [25] | W. Holmes, "Projecting the Need for and Cost of Long-Term Care for Older Persons," 2021. |
| [26] | H. &. W. L. J. •Abdi, "Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.," p. https://doi.org/10.1002/wics.101, 2010. |
| [27] | N. V. G. A. S. A. B. G. K. A. A. D. T. A. E. .. 2. M. 3. Nersisyan S, "exhaustive search-based feature selection for classification and survival regression. PeerJ," p. ;10:e13200. doi: 10.7717/peerj.13200. PMID: , 2022. |
| [28] | T. F. J. T. R. • Hastie, "Linear Methods for Regression. In: The Elements of Statistical Learning. Springer Series in Statistics.," pp. https://doi.org/10.1007/978-0-387-21606-5\_3, 2001. |
| [29] | Z. L. Y. X. L. S. J. W. a. G. -S. X. Z. Zhang, "Discriminative Elastic-Net Regularized Linear Regression," pp. vol. 26, no. 3, pp. 1466-1481 doi: 10.1109/TIP.2017.2651396., 2017. |