Data Analysis

1. Introduction

1.1 What are HAIs and why are they important to study? Give 1-2 examples of common HAIs.

Answer: Healthcare-Associated Infections (HAIs) are infections that happen when people are being treated in hospitals or other healthcare settings. They can occur during medical procedures or from contact with infected surfaces or equipment. Studying HAIs is important because they can make patients sicker, increase medical costs, and sometimes even cause death. By understanding HAIs, we can find ways to prevent them and keep patients safer.

Examples of common HAIs:

I. Central Line-Associated Bloodstream Infections (CLABSIs): These happen when germs get into the bloodstream through a tube inserted into a vein.

II. Catheter-Associated Urinary Tract Infections (CAUTIs): These occur when bacteria enter the urinary tract through a catheter.

1.2 Describe your HAI data set

Answer: The data set is from the Centers for Disease Control and Prevention (CDC). It focuses on Healthcare-Associated Infections (HAIs) in Texas hospitals. The main focus is on CLABSI (Central Line-Associated Bloodstream Infections) and the number of device days. Device days refer to the number of days patients had medical devices, like central lines, in use. This data helps analyze the risk of infections in different hospitals and counties in Texas.

2. Clustering

2.1 For cluster analysis, describe what it is and how it is used. Search the web and find an example of the use of cluster analysis. Discuss that reference and list the reference in APA format at the end of your paper.

Answer: Cluster analysis is a method used to group data points that are similar to each other into clusters. It helps identify patterns in data by dividing it into meaningful categories or groups based on similarities. This technique is often used in fields like marketing, biology, and healthcare to find patterns that may not be obvious at first.

In healthcare, cluster analysis can be used to group patients with similar medical conditions or hospital infection rates. This allows researchers to study specific groups in more detail, leading to better decision-making and improved outcomes.

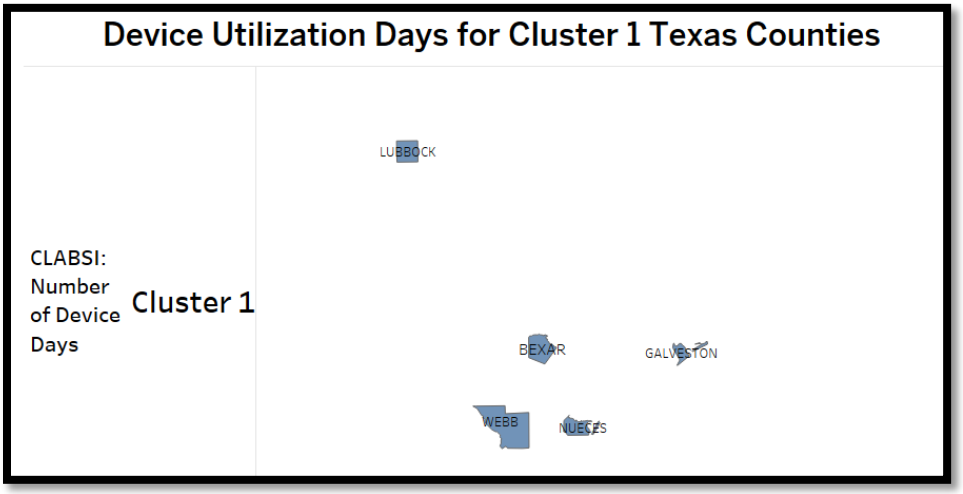
An example of cluster analysis is its use in customer segmentation in marketing. For instance, a company might use cluster analysis to group customers based on their purchasing habits. This helps the company understand which groups of customers are more likely to buy certain products. By identifying these clusters, the company can create targeted marketing strategies for each group, improving sales and customer satisfaction.

Reference: Berkhin, P. (2006). A survey of clustering data mining techniques. *Grouping Multidimensional Data*, 25-71.

https://doi.org/10.1007/3-540-28349-8_2

2.2 Put the appropriate graphs and test results from your cluster analysis in each section and discuss the results.

Answer:



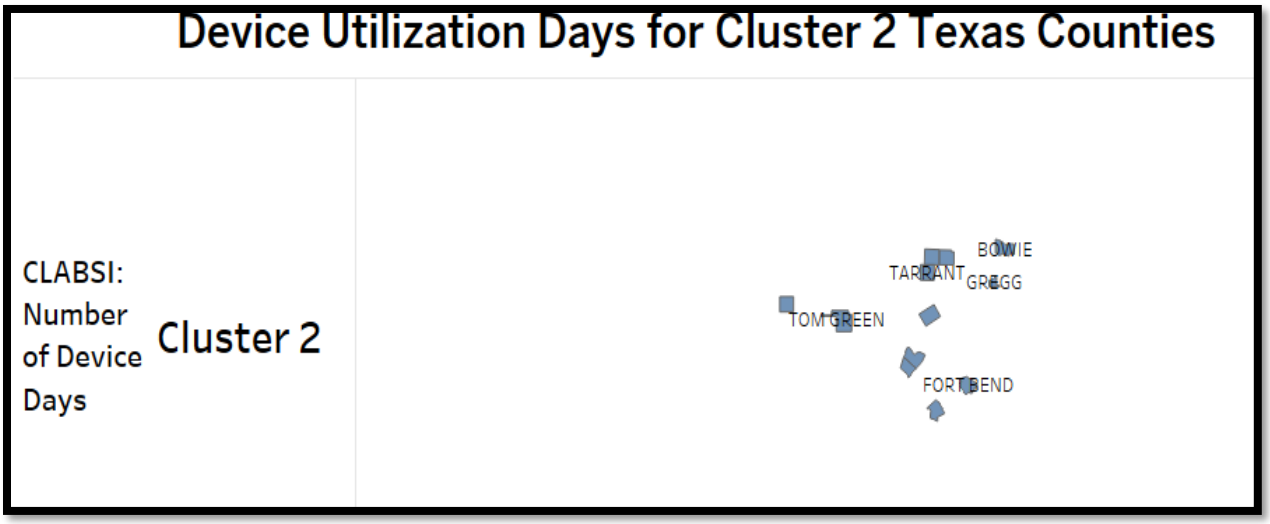
Summary	
Count:	5
SUM(Number of Device Days)	
Sum:	652,741
Average:	108,790.17
Minimum:	20,428
Maximum:	374,999
Median:	40,053.00
Standard deviation:	139,450
Skewness:	1.37
Excess Kurtosis:	0.29

Clusters

Cluster 1

Cluster 1

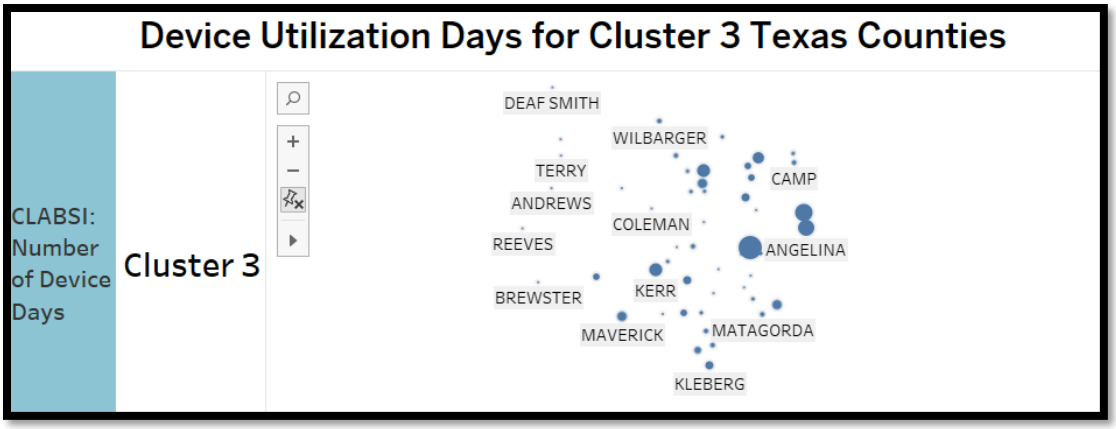
Analysis of Cluster 1: Cluster 1 consists of only 5 hospitals with a significantly high average number of device days (108,790.17). The high standard deviation (139,450) indicates a wide variation in the number of device days across these hospitals. The skewness of 1.37 shows that the data is moderately right-skewed, meaning a few hospitals have much higher device days than the others. The kurtosis (0.29) suggests that the distribution is close to normal but with slightly thinner tails



Summary	
Count:	12(100.0%)
SUM(Number of Device Days)	
Sum:	352,222 (42.6%)
Average:	29,351.83
Minimum:	7,400
Maximum:	115,089
Median:	14,472.50
Standard deviation:	33,591
Skewness:	1.73
Excess Kurtosis:	1.72

Cluster 2

Analysis of Cluster 2: Cluster 2 contains 12 hospitals and shows a lower average (29,351.83) compared to Cluster 1. The standard deviation (33,591) suggests moderate variability in device days between hospitals. With skewness of 1.73, this cluster also exhibits a right-skewed distribution, indicating that some hospitals have significantly higher device days. The kurtosis of 1.72 indicates heavier tails, which means there are more extreme values (outliers) in this cluster.



Summary	
Count:	51
SUM(Number of Device Days)	
Sum:	63,840
Average:	769.16
Minimum:	5
Maximum:	8,877
Median:	298.00
Standard deviation:	1,408
Skewness:	3.61
Excess Kurtosis:	14.83

Cluster 3

Analysis of Cluster 3: Cluster 3 has the largest number of hospitals (51) but the lowest average number of device days (769.16). The data is highly right-skewed (skewness of 3.61), indicating that most hospitals have very few device days, but a few hospitals have much higher numbers. The kurtosis value of 14.83 indicates very heavy tails, meaning there are extreme outliers that significantly affect the overall distribution.

3. Descriptive Statistics

3.1 Define descriptive statistics:

Answer: Descriptive statistics are a set of tools used to summarize and describe the main features of a dataset. They provide simple summaries about the sample and its measures without making any inferences or predictions. Descriptive statistics help simplify large amounts of data in a clear and understandable way. Key elements of descriptive statistics include:

3.1.1 Measures of Central Tendency:

- Mean (Average): The sum of all values divided by the number of values.
- Median: The middle value when the data is ordered from smallest to largest.
- Mode: The most frequently occurring value in the data set.

3.1.2 Measures of Spread:

- ✓ Range: The difference between the highest and lowest values.
- ✓ Variance: Measures of how much the data points differ from the mean.
- ✓ Standard Deviation: A more commonly used measure of the spread of data around the mean.

3.1.3 Shape of Distribution:

- Skewness: Indicates whether the data is symmetric or skewed (right or left).
- Kurtosis: Tells how the tails of the distribution compare to a normal distribution (whether they are more or less extreme).

3.1.4 Descriptive statistics are commonly used to get a quick overview of data before performing further statistical analysis.

3.2 State the descriptive statistics for this data set.

Answer:

Summary	
Count:	117
SUM(Number of Device Days)	
Sum:	1,543,091
Average:	13,188.81
Minimum:	5
Maximum:	374,999
Median:	583.00
Standard deviation:	43,386
Skewness:	6.16
Excess Kurtosis:	43.47

- **Count:** There are 117 counties in Texas included in this dataset, each reporting the number of device days.
- **Sum**: The total number of device days across all counties is 1,543,091. This shows the overall usage of medical devices, like central lines, in hospitals across these counties.
- **Average**: The average number of device days is 13,188.81, but this number may not be very representative of most counties because of the high variability in the data.
- **Minimum and Maximum**: The minimum value is 5 device days, while the maximum is 374,999. This large range indicates that some counties have very low device usage, while others have extremely high usage.
- **Median:** The median is 583, which is much lower than the average. This suggests that more than half of the counties have fewer than 583 device days, but a few counties have very high numbers, pulling the average up.
- **Standard Deviation**: A high standard deviation of 43,386 shows that the data is very spread out, meaning there is a lot of variability between counties.
- **Skewness**: The skewness of 6.16 tells us that the data is highly right-skewed. This means that most counties have relatively low device days, but a few counties have much higher numbers, creating a long tail on the right.
- **Excess Kurtosis**: The kurtosis of 43.47 indicates that there are many extreme outliers in the data. This means that a small number of counties have significantly higher device days compared to the rest

4 Discuss the descriptive statistics.

Answer: The data shows that most counties in Texas have relatively low device usage, but a few counties have very high numbers, which skews the averages. This could suggest that larger, more populated counties or those with major hospitals have much higher device days, while smaller counties have far fewer. The high variability and presence of outliers point to significant differences in healthcare infrastructure and patient volume across the state.

5 Use the internet to look up the meaning of each test (e.g. skewness) then explain what your results mean in terms of the data characteristics. For example, is the data slightly or strongly skewed? Is the skew positive or negative?

Answer:

Here is the explanation of each test and what the results mean in terms of the data characteristics:

I. Skewness: Skewness measures the asymmetry of the data distribution. A skewness value close to 0 indicates a symmetrical distribution. Positive skewness means the data has a long tail on the right (more values are concentrated on the lower end, but a few large values pull the distribution to the right). Negative skewness means the data has a long tail on the left.

The skewness in our data is 6.16, which is a very high positive value. This means the data is strongly right skewed. Most counties have low device days, but a few counties have very high device days, pulling the average upward.

II. Kurtosis: Kurtosis measures the tailed Ness of the data distribution. It shows whether the data has more or fewer extreme values than a normal distribution (kurtosis = 3). High kurtosis means the data has heavy tails (many outliers), while low kurtosis indicates light tails (fewer outliers).

Result: The excess kurtosis in your data is 43.47, which is extremely high. This indicates that the data has a lot of extreme outliers. A few counties have much higher device days than most, making the distribution highly peaked with long tails.

III. What the Results Mean in Terms of Data Characteristics:

- The data is strongly positively skewed, meaning most counties have relatively low device days, while a few have very high values.
- The high kurtosis shows that the data has many extreme outliers, which are counties with significantly higher device days than the rest.
- The large standard deviation confirms that there is a wide spread in the data, with some counties having much higher or lower values.
- The difference between the mean and median further supports that the data is influenced by extreme outliers, with most counties having much lower values than the average.