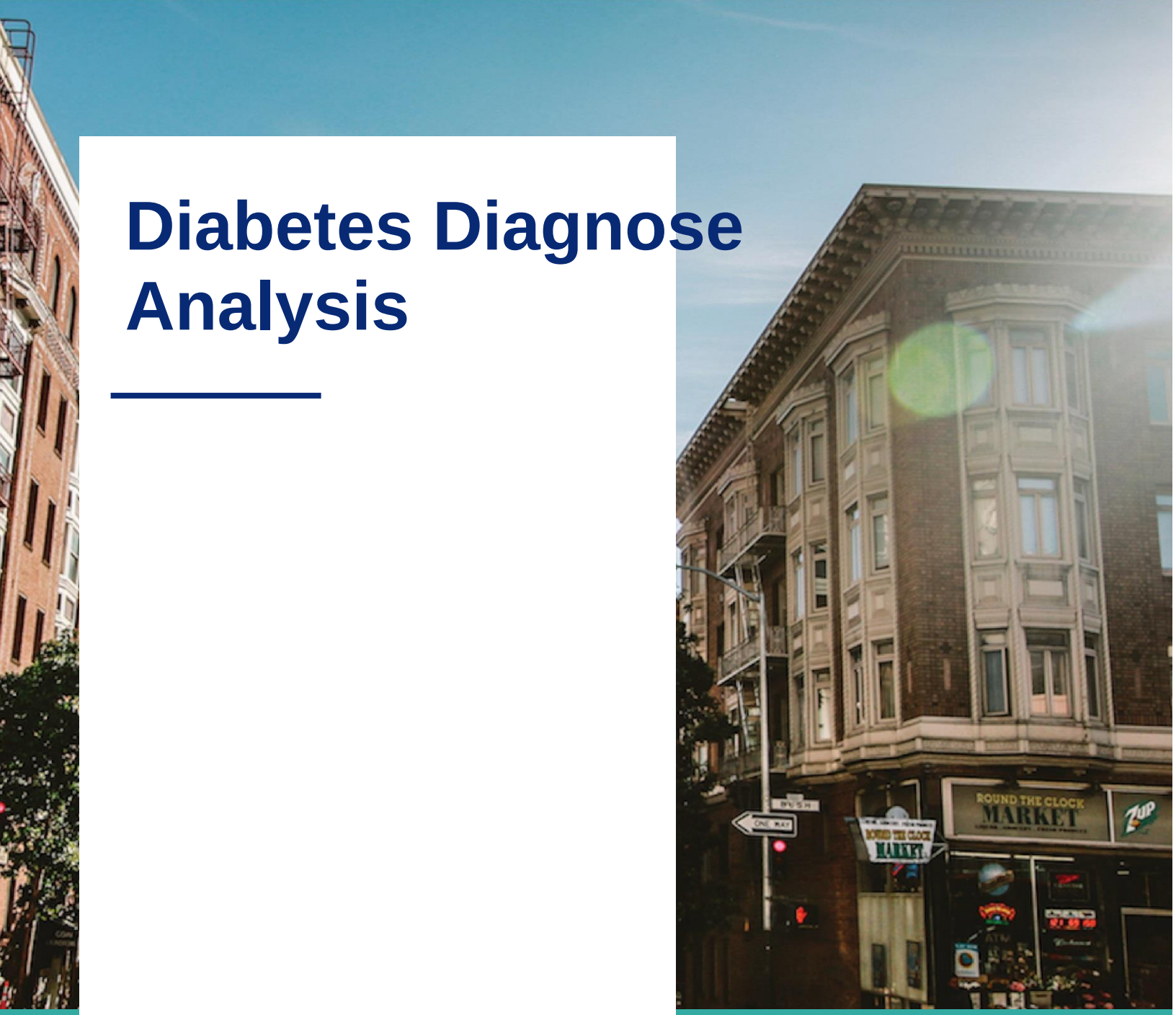
Diabetes Diagnose Analysis



APRIL 29

CS544 A1 Project

Authored by: Fangge You

Table of Contents

Introduction

This project will illustrate the details and provide a better understanding of diabetes diagnose in the United States based on the dataset gathered through kaggle.com[1]. The dataset was originally created by the Centers for Medicare and Medicaid Services, and it was last updated in 2017. This dataset contains total of 163,065 instances and 12 variables. Since the dataset contains the top 100 diagnoses and the information of hospitals that provide those diagnose, only diabetes diagnose was chosen for detailed analysis to create an in-depth understanding of this diagnose.

In the steps of preprocessing the dataset, I first imported the dataset, a csv file, into R Studio. Then, I checked if there is any missing values of any instance. Fortunately, there is no missing value. Therefore, I took out all instances with the DRG Code of 638 which is the code of diabetes diagnose and five other variables (provider's state, total discharges number, average covered charges, average total payments, and average Medicare payments) to form a subset of diabetes. After the preprocessing, there are 1820 instances in the diabetes subset, which means that there are 1820 hospitals in the United States that provide diabetes diagnose (as of Figure 1).

Figure 1: Screenshot of Diabetes Subset

	DRG.Code	Provider.State	Total.Discharges	Average.Covered.Charges	Average.Total.Payments	Average.Medicare.Payments
125167	638	NY	21	$20,006.57	$6,048.85	$5,177.85
127576	638	AL	32	$21,175.81	$4,678.43	$4,047.68
127577	638	AL	12	$9,719.16	$4,863.75	$4,203.41
127578	638	AL	35	$17,021.54	$4,434.57	$3,537.20
127579	638	AL	14	$15,875.50	$5,176.07	$3,394.64
127580	638	AL	12	$13,485.75	$4,222.33	$3,564.66
127581	638	AL	18	$19,038.27	$5,379.66	$3,041.83
127582	638	AL	20	$10,415.45	$4,555.50	$3,654.70
127583	638	AL	17	$15,398.35	$5,554.05	$4,625.58
127584	638	AL	35	$19,273.22	$5,185.65	$3,680.40
127585	638	AL	14	$6,870.57	$4,919.07	$4,200.50
127586	638	AL	31	$6,841.93	$4,628.64	$3,433.32
127587	638	AL	54	$19,911.35	$6,237.29	$5,218.72
127588	638	AL	16	$11,409.93	$4,482.75	$3,569.87

Showing 1 to 15 of 1,820 entries

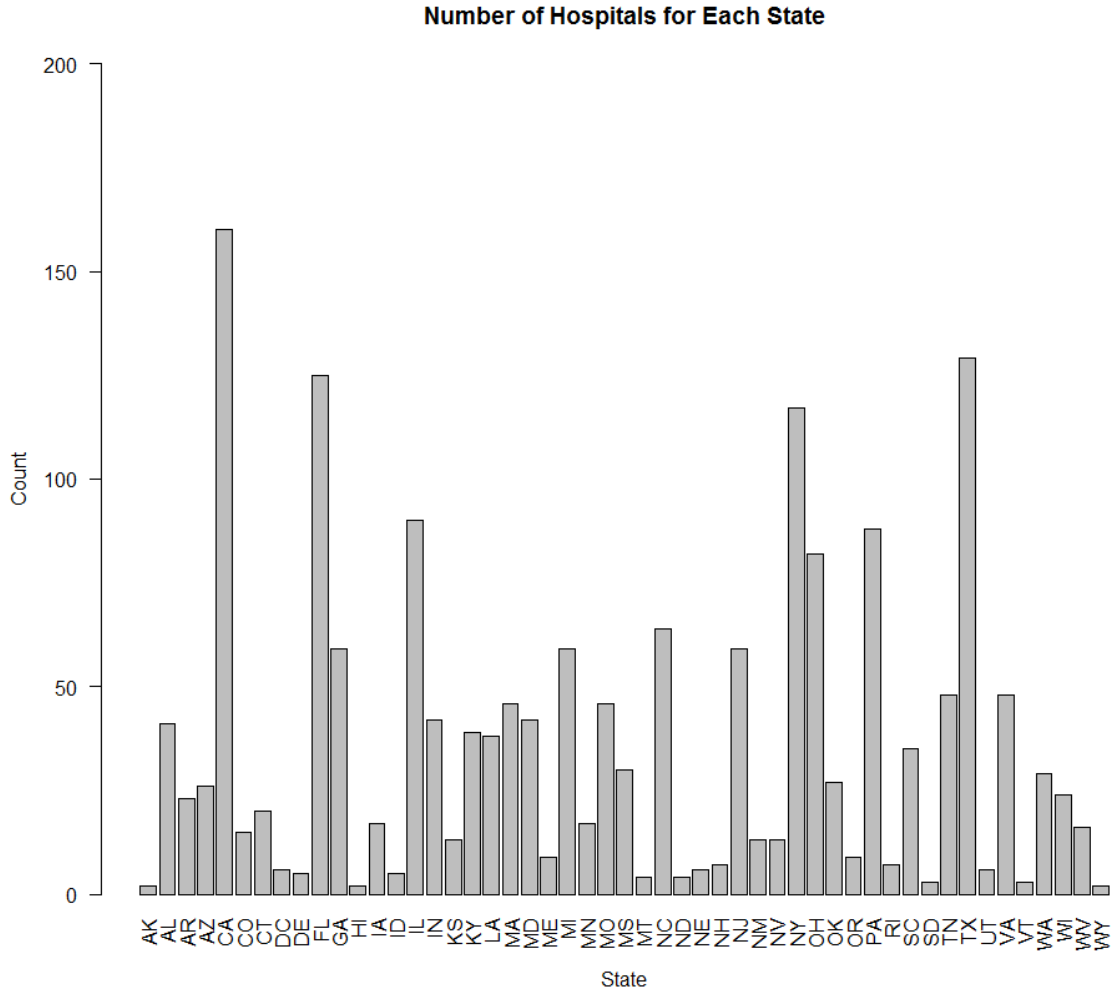
[1] **Dataset Source: https://www.kaggle.com/speedoheck/inpatient-hospital-charges**

Analysis on Categorical and Numerical Data

Categorical Data

The analysis on categorical data was performed on provider's state variable (as of Figure 2). In this analysis, the number of hospitals that provide diabetes diagnose in each state will be illustrated. Based on the results of the analysis, California has 160 hospitals that provide diabetes diagnose, which is the most in the nation. Alaska, Hawaii and Wyoming have only 2 hospitals that provide diabetes diagnose, which is the lowest in the nation. California probably has so many hospitals, or there might be more patients who need diabetes diagnose in California. Compared to larger states, Alaska, Hawaii and Wyoming probably have relatively fewer hospitals, or there are just fewer patients. The reason behind this information still require further research because the dataset does not have the information for concluding the reason behind that.

Figure 2: Barplot of State Variable



Numerical Data

The analysis on numerical data was performed on total discharges number, average total payments, average covered charges, and average Medicare payments. In the dataset, the columns of average total payments, average covered charges, and average Medicare payments contain dollar signs in the values. However, I changed these values to numeric values for more meaningful analysis.

In the analysis on total discharges number, the average number of discharges is 26 among all hospitals that provide diabetes diagnose (as of Figure 3). According to the results, most hospitals have number of discharges under 32. However, the outliers have the number of discharges greater than 57.5 or less than -10.5 (as of Figure 4). Since the number of discharges would never be lower than 0, outliers in this case would have discharges number greater than 57.5. As the results indicate, there are many outliers which have larger discharges number.

Figure 3: Bar Chart of Total Discharges Variable

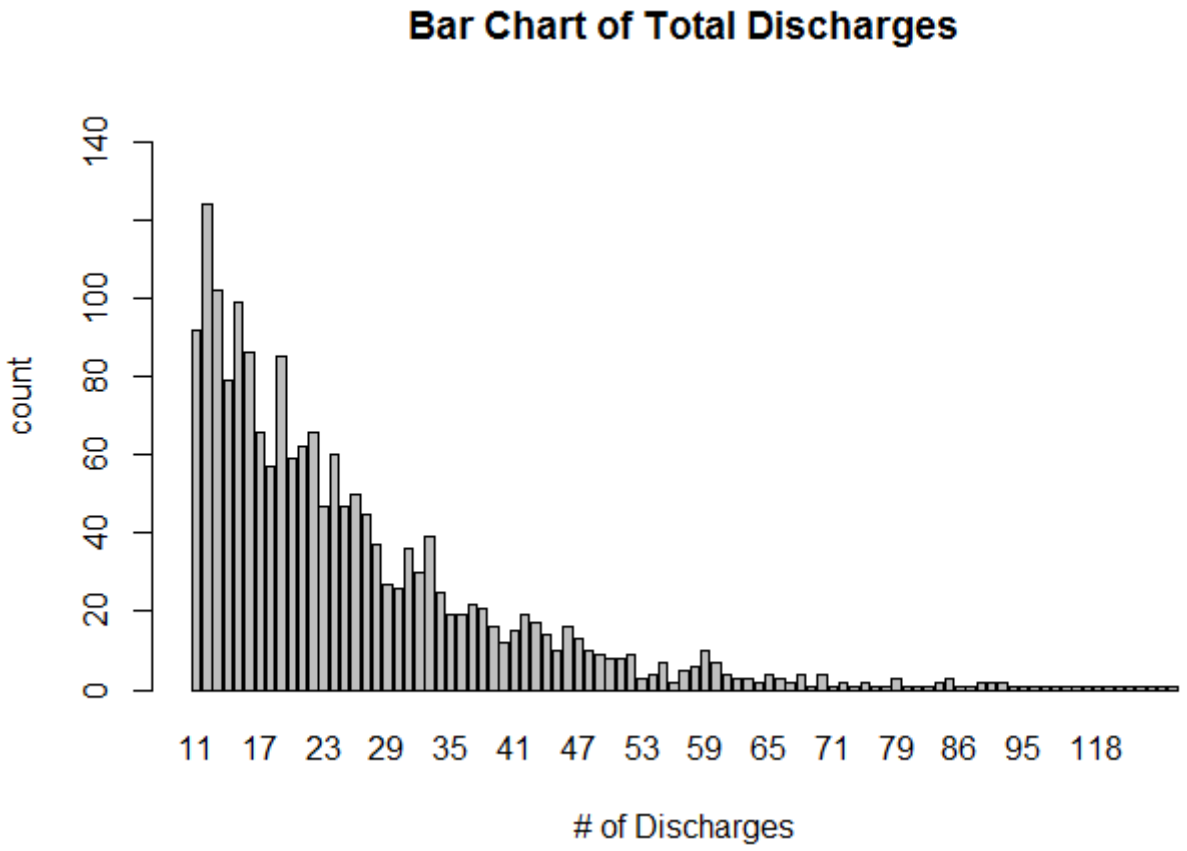
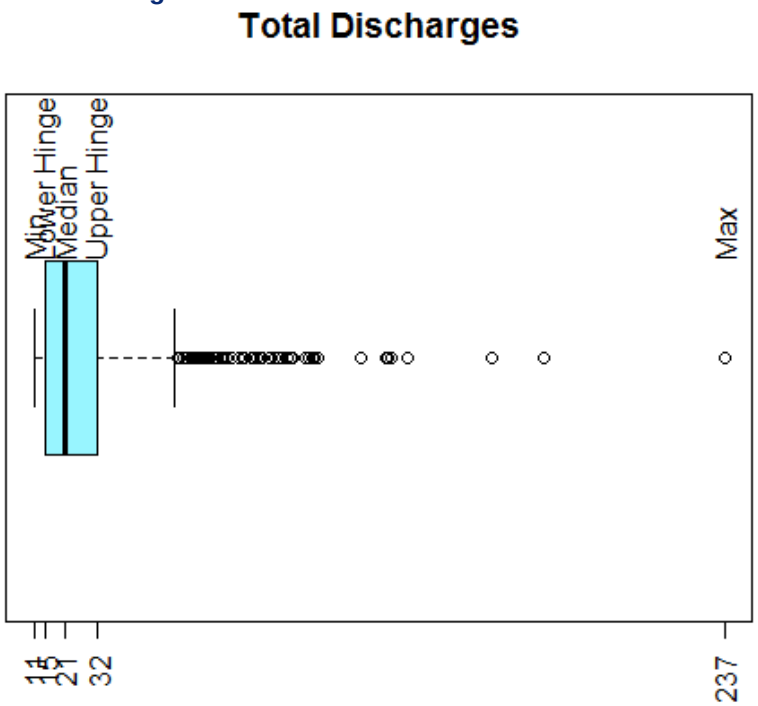


Figure 4: Boxplot of Total Discharges



In the analysis on average total payments, the average total payments of diabetes diagnose in the United States is $89,394 which is the average amount that hospitals would charge a patient on diabetes diagnose. The bar chart shows that the average total payments of diabetes diagnose fluctuate reasonably (as of Figure 5). According to the results, the outliers have average total payments below $29,990.50 or greater than $147,890.50 (as of Figure 6).

Figure 5: Bar Chart of Average Total Payments

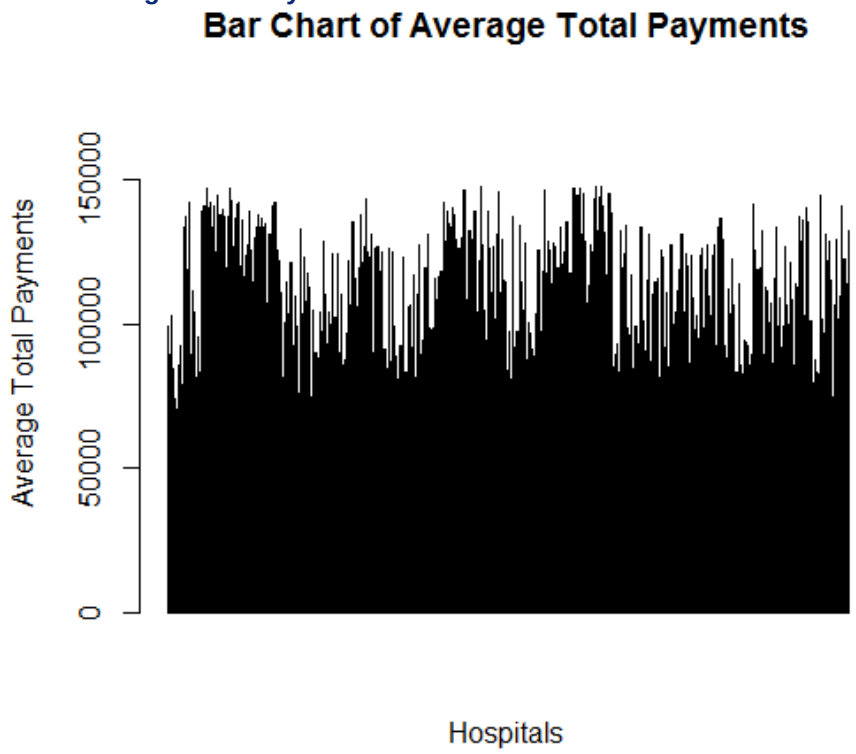
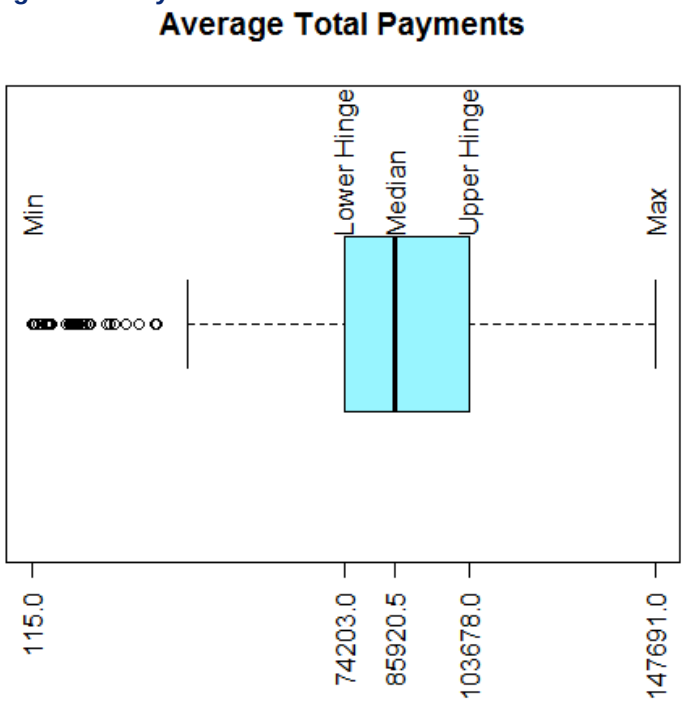


Figure 6: Boxplot of Average Total Payments



In the analysis on average covered charges, the average covered charges of this diagnose is $65,800. The bar chart indicates that the average covered charges of diabetes diagnose fluctuate heavily among different hospitals; the reason of this circumstance is probably due to different Medicare or insurance plans that cover different amount of charges (as of Figure 7). Based on the results, outliers would have average covered charges below -$67,819.75 or greater than $195,514.25. In this case, all instances fall into the reasonable range of average covered charges, and there is no outlier (as of Figure 8). A scatterplot was performed to find any hidden relationship between average total payments and average cover charges (as of Figure 9). In the scatterplot, “tp” (y-axis) stands for average total payments, and “cov” (x-axis) stands for average covered charges. In this case, there is no special relationship between these two variables.

Figure 7: Bar Chart of Average Covered Charges

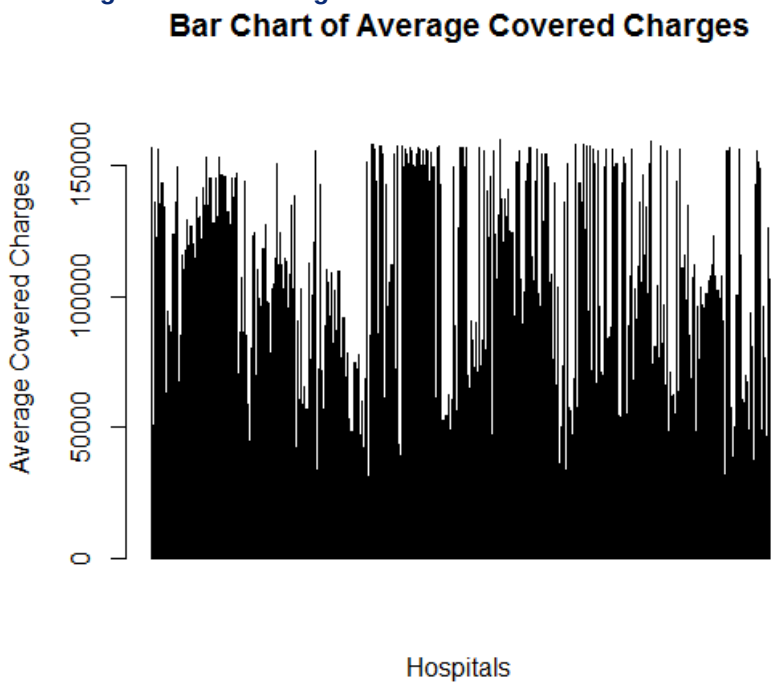


Figure 8: Boxplot of Average Covered Charges

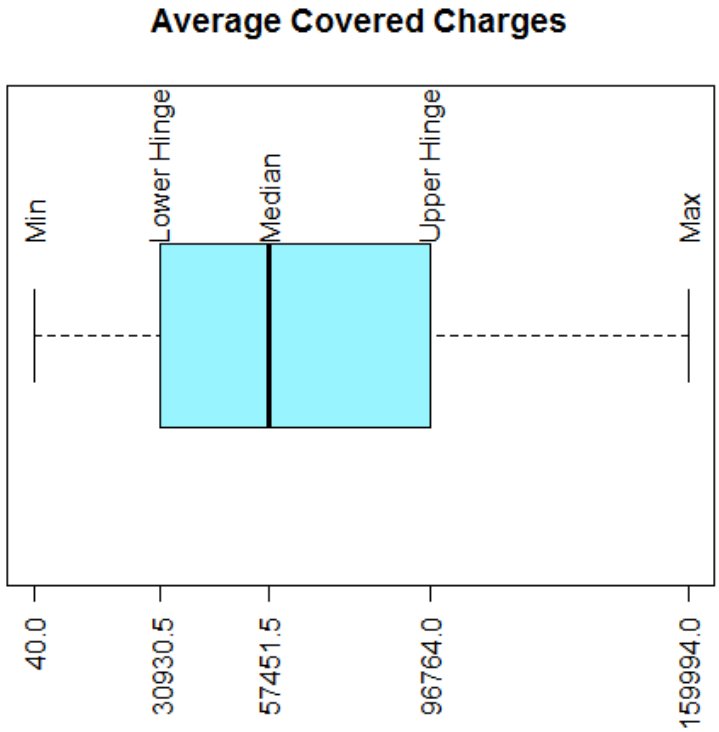
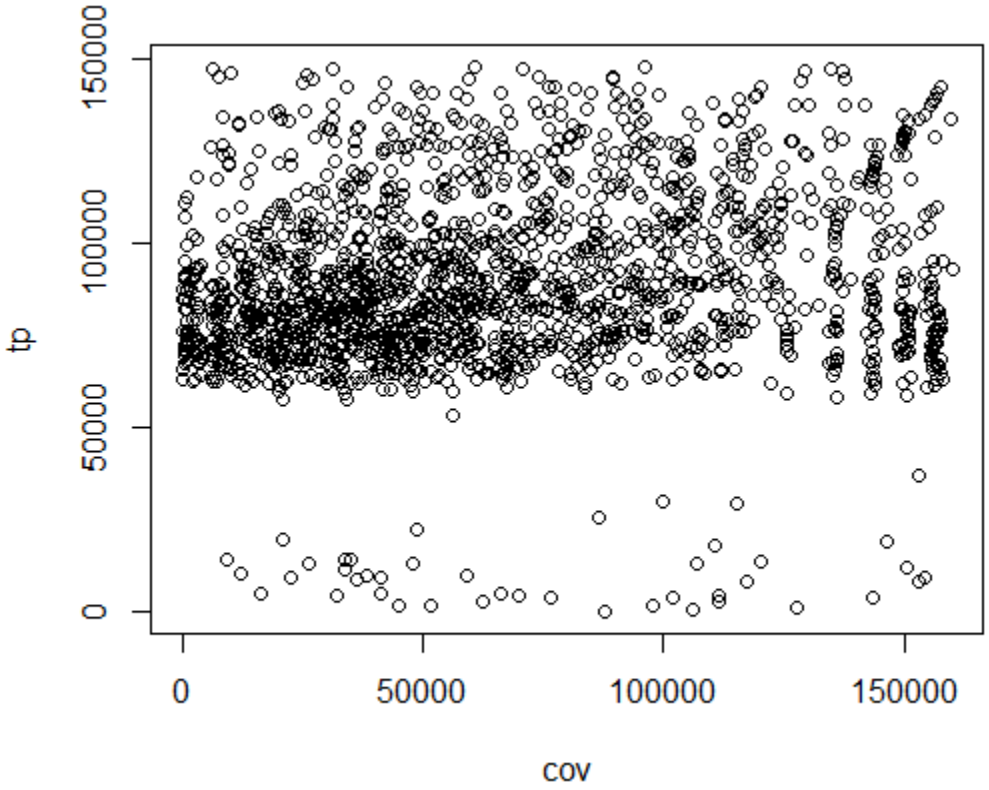


Figure 9: Scatterplot of Average Total Payments & Average Covered Charges



In the analysis on average Medicare payments, the average Medicare payments of this diagnose is $83,066. The bar chart indicates that the average Medicare payments fluctuate reasonably (as of Figure 10). The results indicate that outliers have average Medicare payments below $17,998.25 or greater than $145,528.25 (as of Figure 11). Additionally, a scatterplot between average total payments and average Medicare payments was constructed (as of Figure 12). Within the scatterplot, “tp” represents average total payments, and “med” represents average covered charges. In this case, as the amount of Medicare payments increases, total payments will increase; according to the scatterplot, this relationship is pretty strong because most data fall along a straight line showing a linear increase in the payments.

Figure 10: Bar Chart of Average Medicare Payments

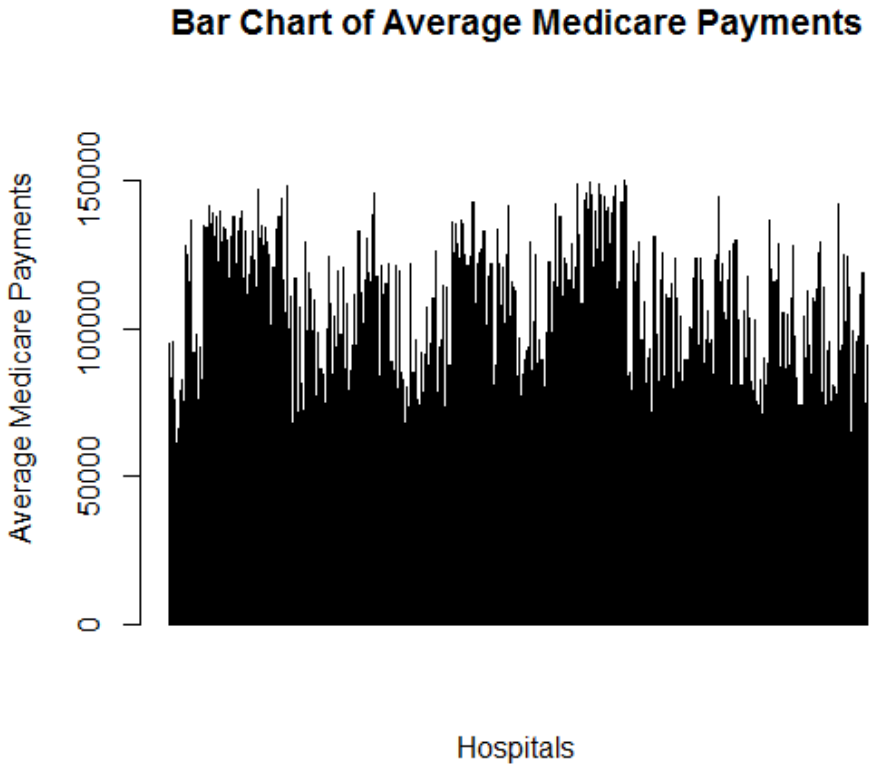


Figure 11: Boxplot of Average Medicare Payments

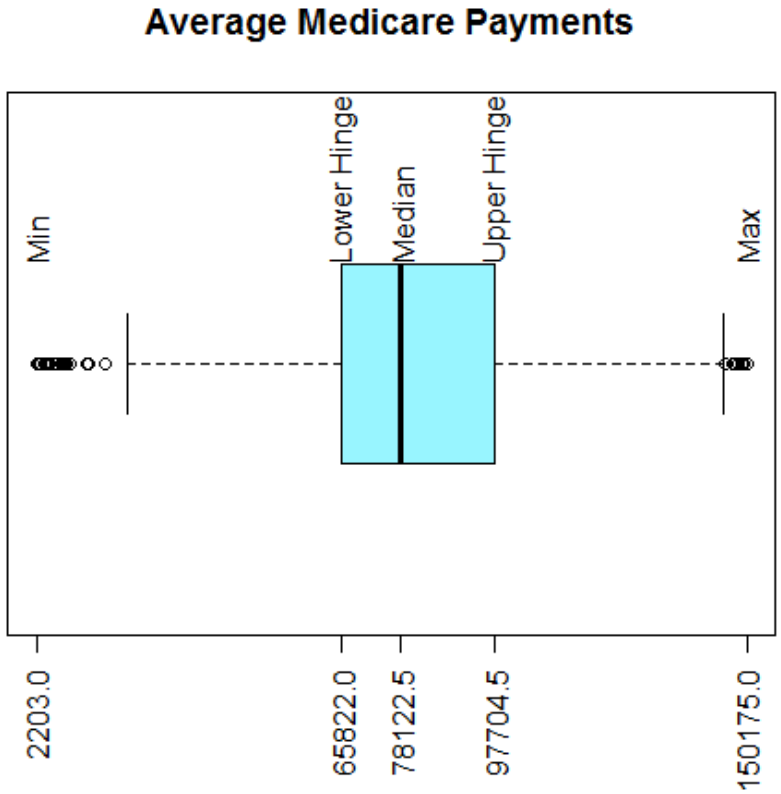
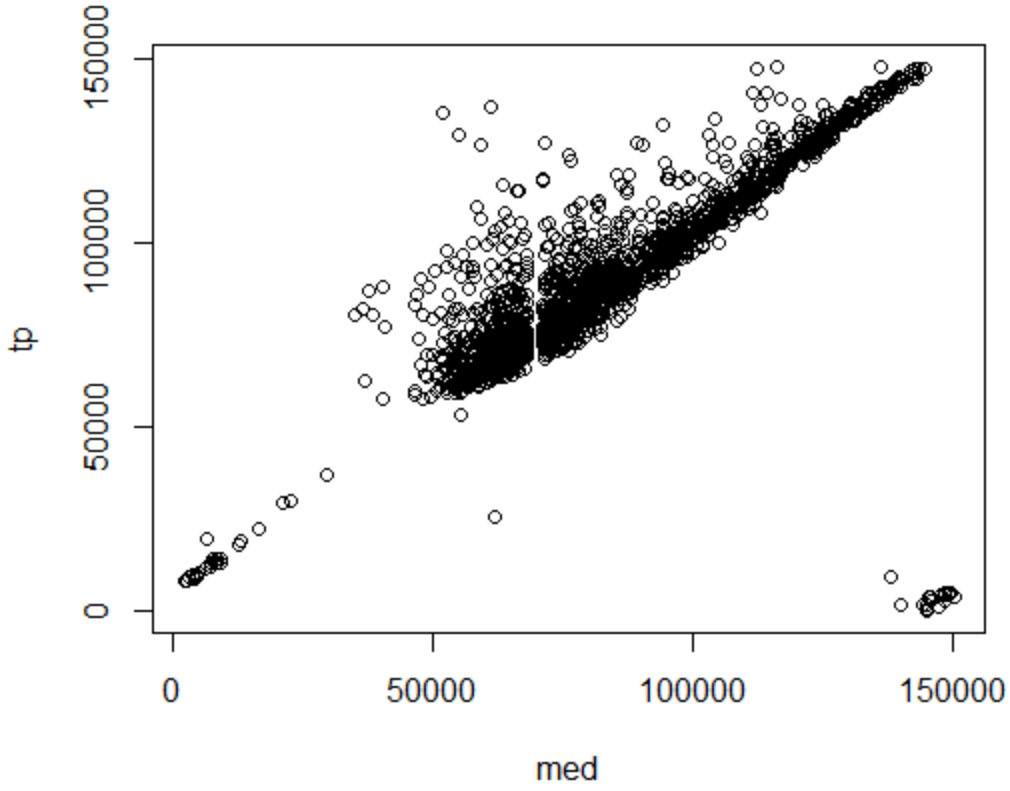


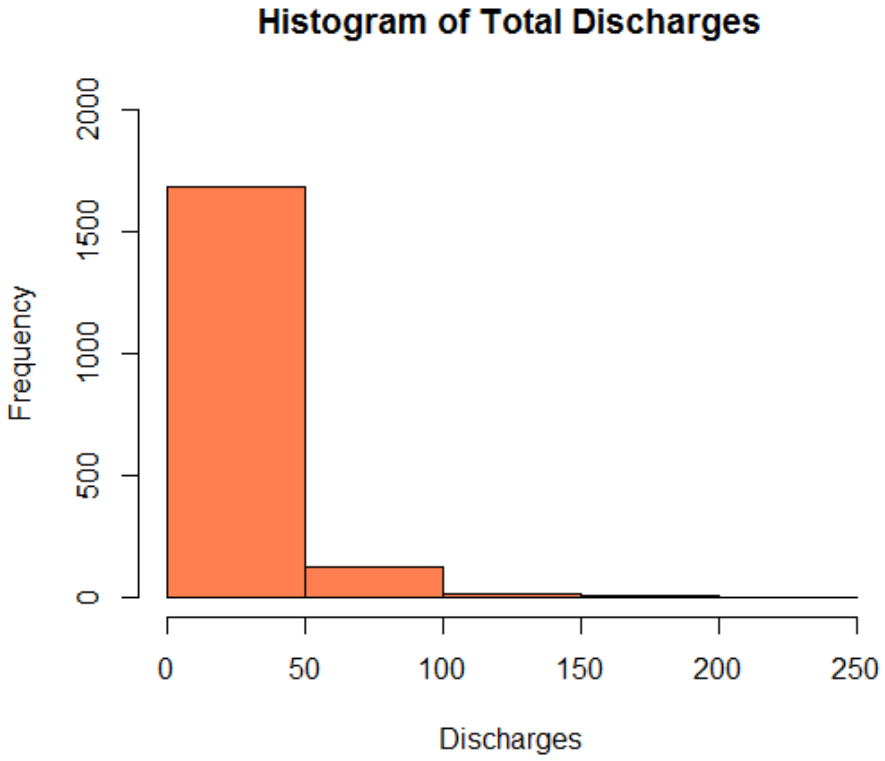
Figure 12: Scatterplot of Average Total Payments & Average Medicare Payments



Distribution of Variables

Discharges number is chosen to be analyzed for further understanding through its distribution. The distribution result indicates that 1686 hospitals have discharges number between 0 and 50, 122 hospitals with discharges number between 50 and 100, 9 hospitals with discharges number between 100 and 150, 2 hospitals with discharges number between 150 and 200, and only 1 hospital with discharges number between 200 and 250. The distribution is skewed to right (as of Figure 13). Most hospitals have discharges number under 50, and only a few have discharges number over 100. In a word, most hospitals in the United States have less than 50 patients discharged from the diabetes diagnose.

Figure 13: Distribution of Total Discharges



Central Limit Theorem

Although the original population is not normally distributed (as of Figure 13), the distribution of samples randomly taken from a given sample size will follow a normal distribution (as of Figure 14). In this case, 1000 random samples of sample size 10, 20, 30, and 40 were selected; in the histograms, besides the distribution of sample size 10 which seems to be a little skewed to right, the other three show a normal distribution. As the sample size increases, the spread of the distribution becomes narrower because standard deviation becomes smaller (as of Figure 15). The population mean is 26.38736, which means that the average total discharges of all instances in diabetes subset is around 26.

Figure 14: Central Limit Theorem

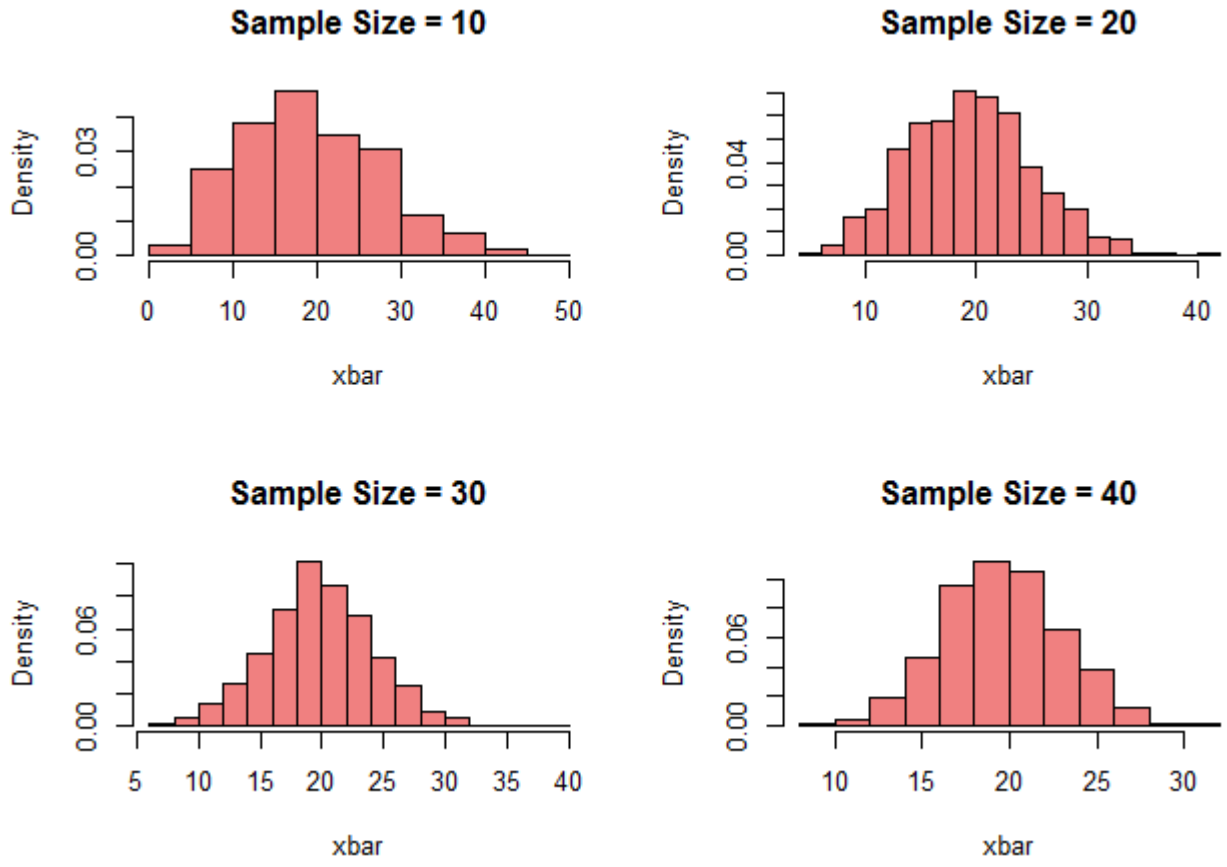


Figure 15: Mean and Standard Deviation

```
Sample Size =  10  Mean =  19.3698  SD =  8.218382
Sample Size =  20  Mean =  19.53805  SD =  5.533554
Sample Size =  30  Mean =  19.7986  SD =  4.30506
Sample Size =  40  Mean =  19.58405  SD =  3.402657
```

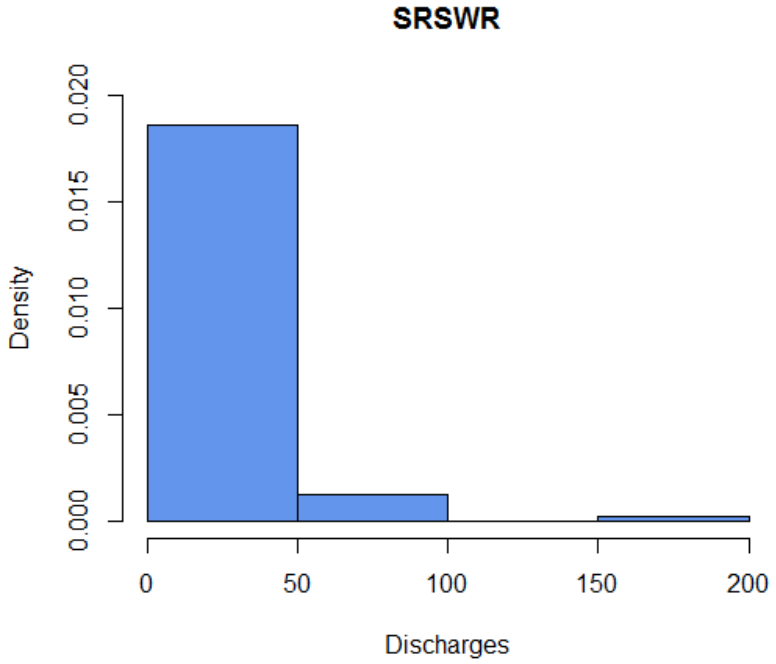
Sampling Methods

In this section, three sampling techniques were applied to the total discharges variable: simple random sampling with replacement, simple random sampling without replacement, and symmetric sampling.

Simple Random Sampling With Replacement

In this sampling method, every sample was randomly picked with replacement from the population. This distribution looks similar to the original distribution of the entire population, and both of them are skewed to right (as of Figure 16).

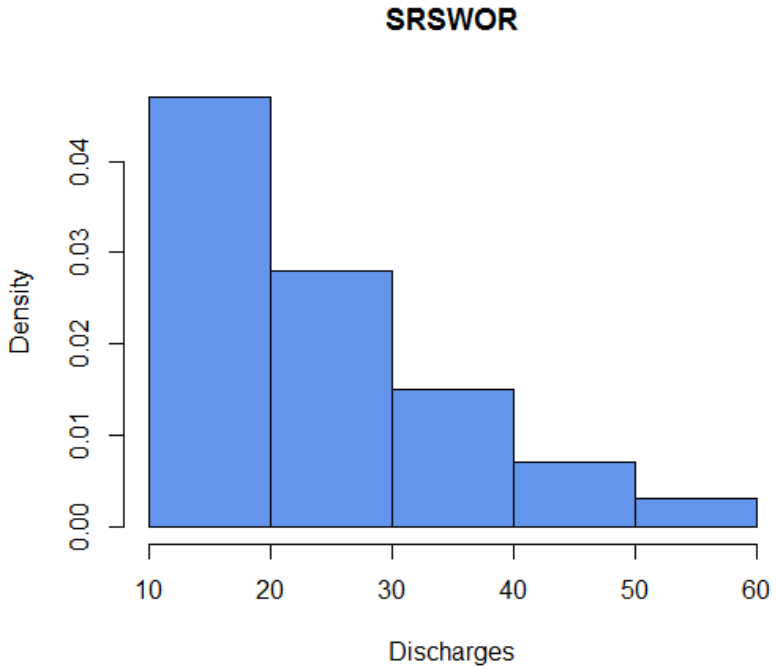
Figure 16: Simple Random Sampling With Replacement



Simple Random Sampling Without Replacement

In this sampling method, every sample was randomly picked without replacement from the population. This distribution is also skewed to right as the original distribution, but it seems to be more perfectly skewed than the original one (as of Figure 17). The reason is probably that as each sample is selected, it is not replaced by another sample in the population before selecting next sample. Therefore, the population size is changing while selecting samples.

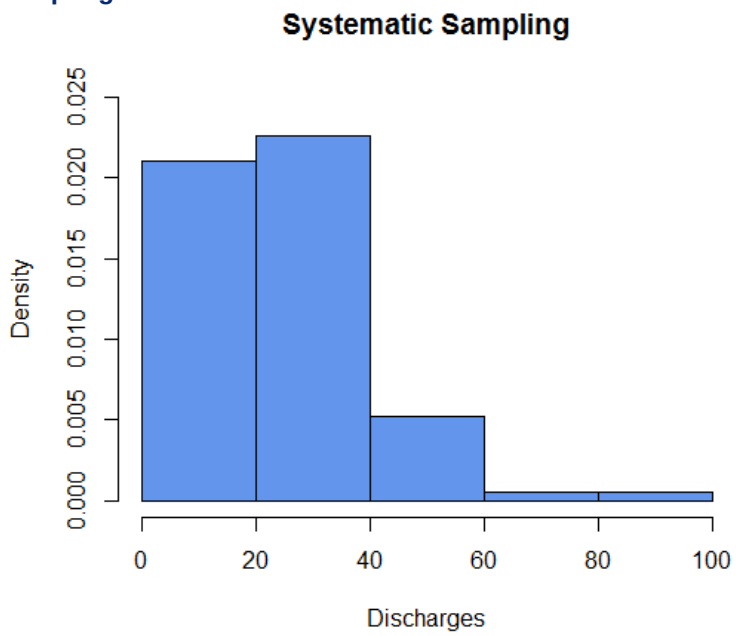
Figure 17: Simple Random Sampling Without Replacement



Systematic Sampling

In this sampling method, every sample was selected through a fixed periodic interval. A sample size of 100 is selected. The first item is randomly selected from the first set of 19 items in the frame. After the first selection, the remaining items are selected by taking every 19th item from the frame. Compared with the population's distribution, this distribution is also skewed to right but with little difference on the left side of the distribution.

Figure 18: Systematic Sampling



Confidence Intervals

For the confidence levels of 80 and 90, the confidence intervals of population and different samples are conducted (as of Figure 19, 20, 21, 22).

Figure 19: Confidence Interval of Population

```
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), t: %.2f , %.2f",
+                  100*(1-i), i,
+                  qt(i/2, df = n-1),
+                  qt(1 - i/2, df = n-1))
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), t: -1.29 , 1.29
90% Conf Level (alpha = 0.10), t: -1.66 , 1.66
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), CI = %.2f - %.2f",
+                  100*(1-i), i,
+                  xbar - qt(1 - i/2, df = n-1) * sd.sample.means,
+                  xbar + qt(1 - i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), CI = 22.44 - 25.76
90% Conf Level (alpha = 0.10), CI = 21.96 - 26.24
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), Precision = %.2f",
+                  100*(1-i), i,
+                  2* qt(1-i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), Precision = 3.33
90% Conf Level (alpha = 0.10), Precision = 4.28
```

Figure 20: Confidence Interval of SRSWR

```
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), t: %.2f , %.2f",
+                  100*(1-i), i,
+                  qt(i/2, df = n-1),
+                  qt(1 - i/2, df = n-1))
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), t: -1.29 , 1.29
90% Conf Level (alpha = 0.10), t: -1.66 , 1.66
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), CI = %.2f - %.2f",
+                  100*(1-i), i,
+                  xbar - qt(1 - i/2, df = n-1) * sd.sample.means,
+                  xbar + qt(1 - i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), CI = 23.99 - 29.41
90% Conf Level (alpha = 0.10), CI = 23.21 - 30.19
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), Precision = %.2f",
+                  100*(1-i), i,
+                  2* qt(1-i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), Precision = 5.43
90% Conf Level (alpha = 0.10), Precision = 6.98
```

Figure 21: Confidence Interval of SRSWOR

```
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), t: %.2f , %.2f",
+                  100*(1-i), i,
+                  qt(i/2, df = n-1),
+                  qt(1 - i/2, df = n-1))
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), t: -1.29 , 1.29
90% Conf Level (alpha = 0.10), t: -1.66 , 1.66
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), CI = %.2f - %.2f",
+                  100*(1-i), i,
+                  xbar - qt(1 - i/2, df = n-1) * sd.sample.means,
+                  xbar + qt(1 - i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), CI = 22.43 - 25.29
90% Conf Level (alpha = 0.10), CI = 22.02 - 25.70
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), Precision = %.2f",
+                  100*(1-i), i,
+                  2* qt(1-i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), Precision = 2.86
90% Conf Level (alpha = 0.10), Precision = 3.68
```

Figure 22: Confidence Interval of Systematic Sampling

```
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), t: %.2f , %.2f",
+                  100*(1-i), i,
+                  qt(i/2, df = n-1),
+                  qt(1 - i/2, df = n-1))
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), t: -1.31 , 1.31
90% Conf Level (alpha = 0.10), t: -1.70 , 1.70
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), CI = %.2f - %.2f",
+                  100*(1-i), i,
+                  xbar - qt(1 - i/2, df = n-1) * sd.sample.means,
+                  xbar + qt(1 - i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), CI = 24.82 - 33.18
90% Conf Level (alpha = 0.10), CI = 23.58 - 34.42
> for (i in alpha) {
+   str <- sprintf("%2d%% Conf Level (alpha = %.2f), Precision = %.2f",
+                  100*(1-i), i,
+                  2* qt(1-i/2, df = n-1) * sd.sample.means)
+   cat(str,"\n")
+ }
80% Conf Level (alpha = 0.20), Precision = 8.36
90% Conf Level (alpha = 0.10), Precision = 10.83
```

Conclusion

This project has provided in-depth understanding of diabetes diagnose in the United States. By taking closer look at different variables in the dataset, a comprehensive view of diabetes diagnose in the nation is depicted. Although payments and number of discharges are different for each hospital and individual patient's condition, choosing which hospital for this diagnose is highly dependent on individuals. For more detailed understanding of this diagnose in healthcare industry, further analysis and research will be needed. Most importantly, it is more critical to avoid paying huge amount of money to receive any medical diagnose by keeping healthy. Wish everyone good health!