Provider based analysis of Hospice Care

https://github.com/kcw2144-ps3060/EDAV

1. Introduction

Between 1975 and 2016, life expectancy at birth increased from 72.6 to 78.8 years for the total U.S. population (https://www.cdc.gov/nchs/data/hus/hus16.pdf#014), and it is expected to rise up to 85 years by 2050 (https://www.census.gov/prod/2014pubs/p25-1140.pdf). However, healthcare system, designed in 1965 to care for such increasing aging population, has not kept pace and aligned with the needs. Also, terminal illnesses create significant expenses for the Medicare program and present a major disadvantage to the quality of life. Hospice care was designed to provide a better quality of life to the patients and have less burden on the Medicare system. Hospice care is given to patients who are terminally ill with a life expectancy of 6 months or less.With the available hospice care data for the years 2016 - 2014, we present trends in hospice care and identify hospice care and service utilization. Identification of potential issues and service utilization will benefit the entire Medicare system and on the terminally ill population. Group responsibilities are listed as follow: Member 1.(Kushal :kcw2144)-data preprocessing, analysis, and visualization; Member 2.(Paridhi:ps3060)-data preprocessing, interactive component designing and implementation.

2. Description of data

Center for Medicare & Medicaid Services (CMS's) released Pubic User Files (PUF) (https: //www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/ Medicare-Provider-Charge-Data/Hospice2016.html), for the Hospice for fiscal year 2016, which provides information on services provided to Medicare beneficiaries by hospice providers. This PUF is based on information from CMS's Chronic Conditions Data Warehouse data files and the spending and utilization data is aggregated to the hospice provider. The population of our study is all Medicare beneficiaries who submit at least of one day of Hospice services for the particular fiscal year.

2.1 Data Cleaning

The Hospice PUF variables are checked for potentially miscoded entries for categorical variables and outliers for continuous variables.We used a seperate R script to do data cleaning(R script: https://github.com/kcw2144-ps3060/EDAV/blob/master/Healthcare_ Provider_Data_Preprocess.Rmd, https://github.com/kcw2144-ps3060/EDAV/blob/master/ Days/EDAV_Days.Rmd)

3. Analysis of data quality

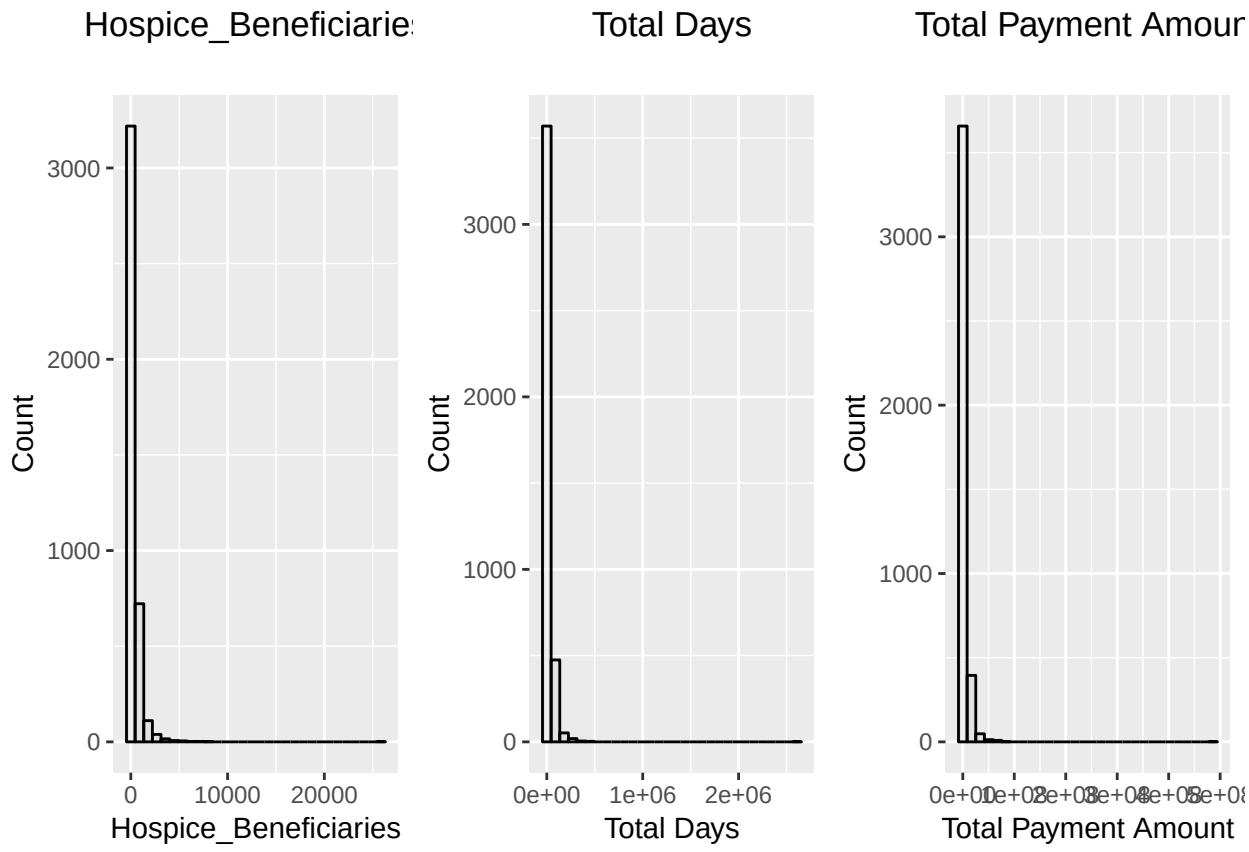
The Hospice PUF contains information on hospice utilization, payment, submitted charges, primary diagnoses, sites of service, type of care services, length of services and hospice beneficiary demographics. There were 7 categorical variables and 40 continuous variables in our data set.

3.1 General Distribution

General trend in our data distribution is depicted in following figures where the variables are right skwed.

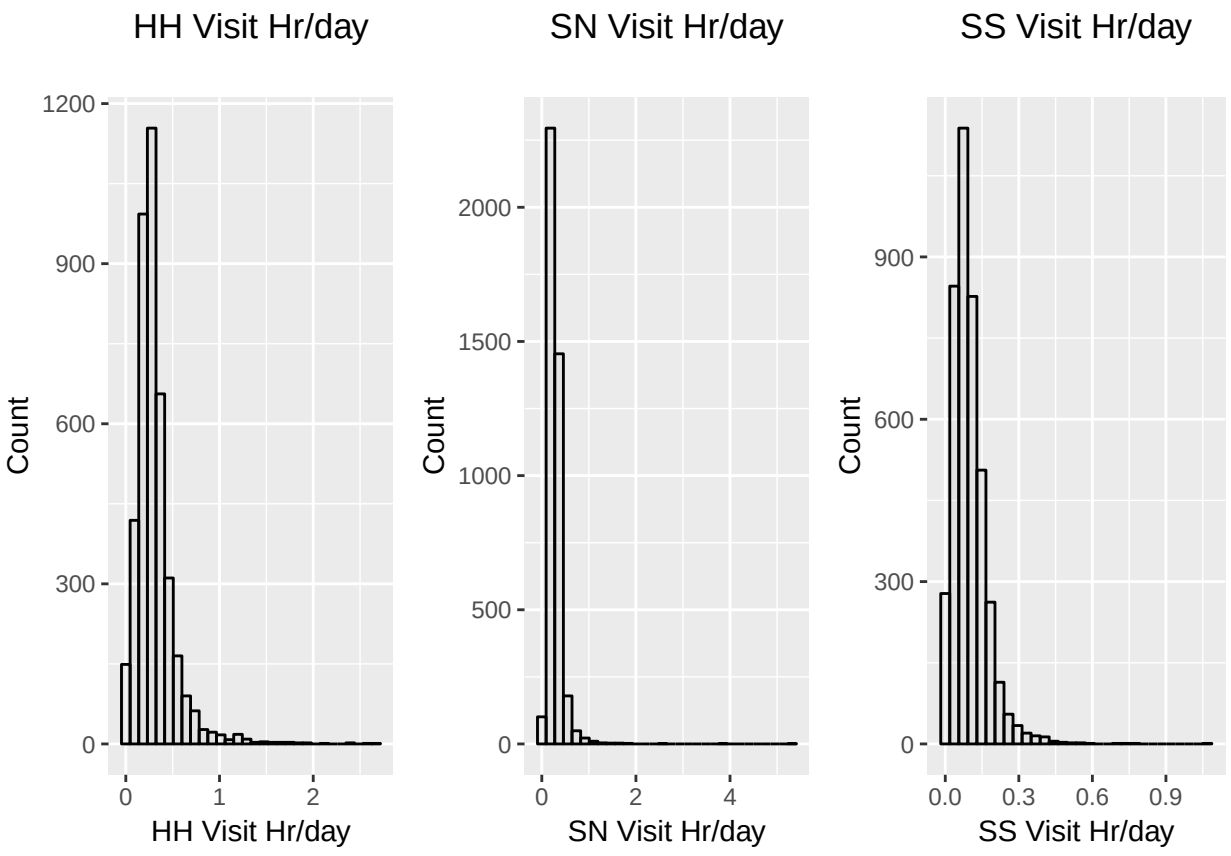
```
patientcare.data.2016C <- read.csv("./Healthcare_Provider_2016C.csv",
                                   stringsAsFactors = FALSE,
```

```
                                    na.strings = NA )
patientcare.data.2016C <- (patientcare.data.2016C[, -1])
patientcare.data.2016C[, 1] <- as.character(patientcare.data.2016C[, 1])
patientcare.data.2016C[c(6,8:47)] <- lapply(patientcare.data.2016C[c(6,8:47)],
                                            as.numeric)
patientcare.data.2016C <- as.data.table(patientcare.data.2016C)


d <- ggplot(data = patientcare.data.2016C, aes(Hospice_beneficiaries)) +
    geom_histogram(col = "black", fill = "grey", alpha = 0.2) +
    labs(title = "Hospice_Beneficiaries
") +     labs(x = "Hospice_Beneficiaries", y = "Count") + theme(plot.title = element_text(hjust = 0.5))

e <- ggplot(data=patientcare.data.2016C, aes(Total_Days
)) + geom_histogram( col="black", fill="grey", alpha = .2) +   labs(title="Total Days
") + labs(x="Total Days", y="Count") +  theme(plot.title = element_text(hjust = 0.5))

f <- ggplot(data=patientcare.data.2016C, aes(Total_Medicare
)) + geom_histogram( col="black", fill="grey", alpha = .2) +   labs(title="Total Payment Amount
") + labs(x="Total Payment Amount", y="Count") +  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(d,e,f, ncol=3, nrow= 1, widths=c(20,20,20))
```



Distribution of Care services data

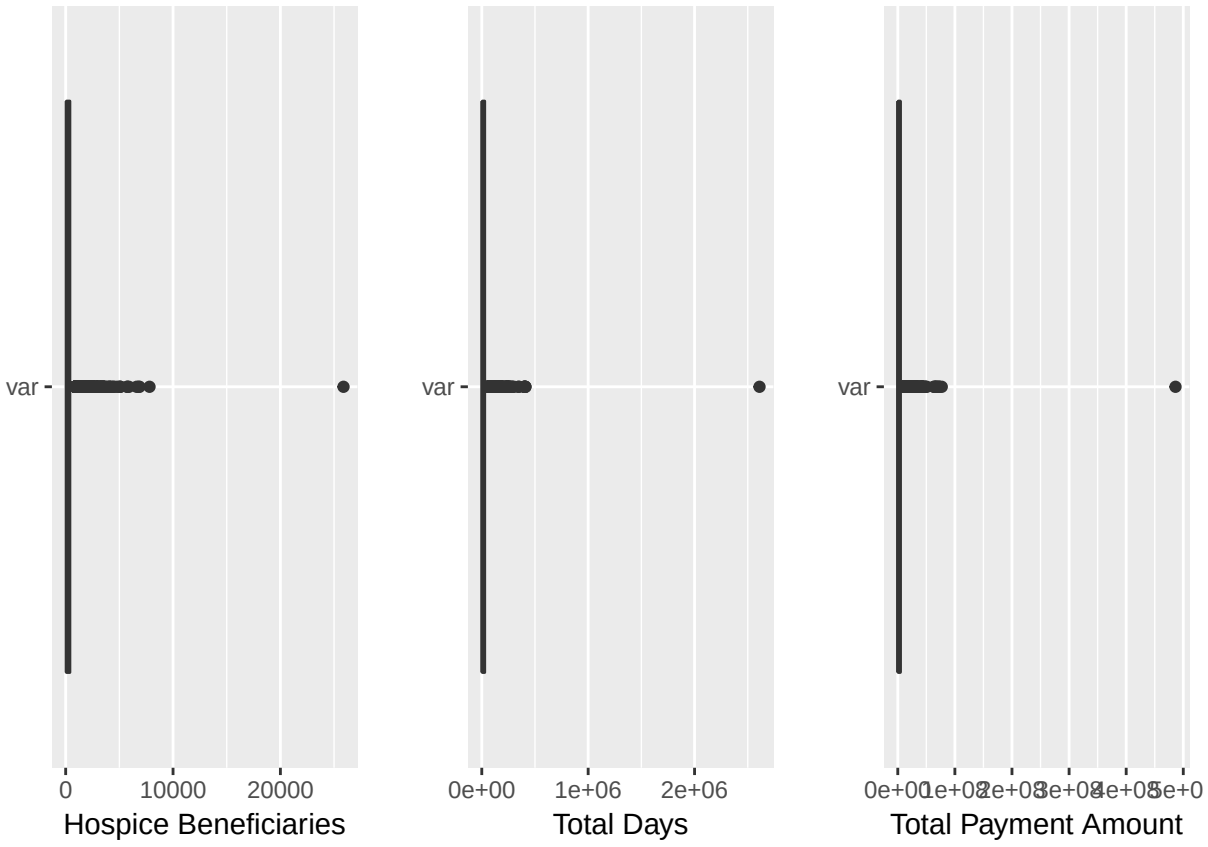
```
library(gridExtra)

a <- ggplot(data=patientcare.data.2016C, aes(Home_Health_Hrs.D
)) + geom_histogram( col="black", fill="grey", alpha = .2) +   labs(title="HH Visit Hr/day
") + labs(x="HH Visit Hr/day", y="Count") +  theme(plot.title = element_text(hjust = 0.5))

b <- ggplot(data=patientcare.data.2016C, aes(Skilled_Nursing_Hrs.D
)) + geom_histogram( col="black", fill="grey", alpha = .2) +   labs(title="SN Visit Hr/day
") + labs(x="SN Visit Hr/day", y="Count") +  theme(plot.title = element_text(hjust = 0.5))

c <- ggplot(data=patientcare.data.2016C, aes(Social_Service_Hrs.D_LW
)) + geom_histogram( col="black", fill="grey", alpha = .2) +   labs(title="SS Visit Hr/day
") + labs(x="SS Visit Hr/day", y="Count") +  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(a,b,c, ncol=3, nrow= 1, widths=c(20,20,20))
```



From the histogram we can see most of our data are right-skewed.

Checking outliers in ,Hospice_beneficiaries, Total_Days, and Total_Medicare_Payment_Amount

```
aa <- ggplot(patientcare.data.2016C,aes("var",Hospice_beneficiaries)) + geom_boxplot() + xlab("") + ylab
bb <- ggplot(patientcare.data.2016C,aes("var",Total_Days)) + geom_boxplot() + xlab("")  + ylab("Total Da
cc <- ggplot(patientcare.data.2016C,aes("var",Total_Medicare)) + geom_boxplot() + xlab("") + ylab("Total
grid.arrange(aa,bb,cc, ncol=3, nrow= 1, widths=c(20,20,20))
```



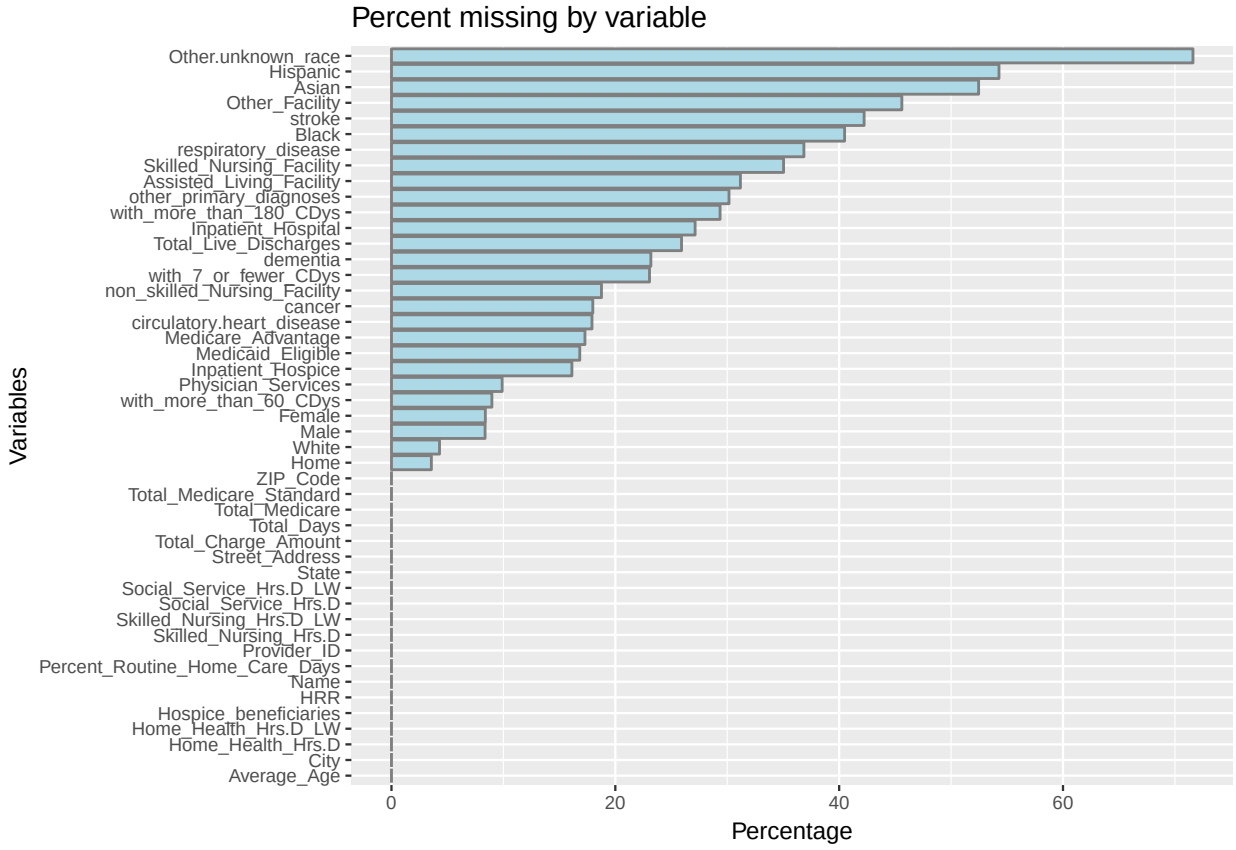
From the box plot we can see some outliers in the our variable.

3.2 Missing Data

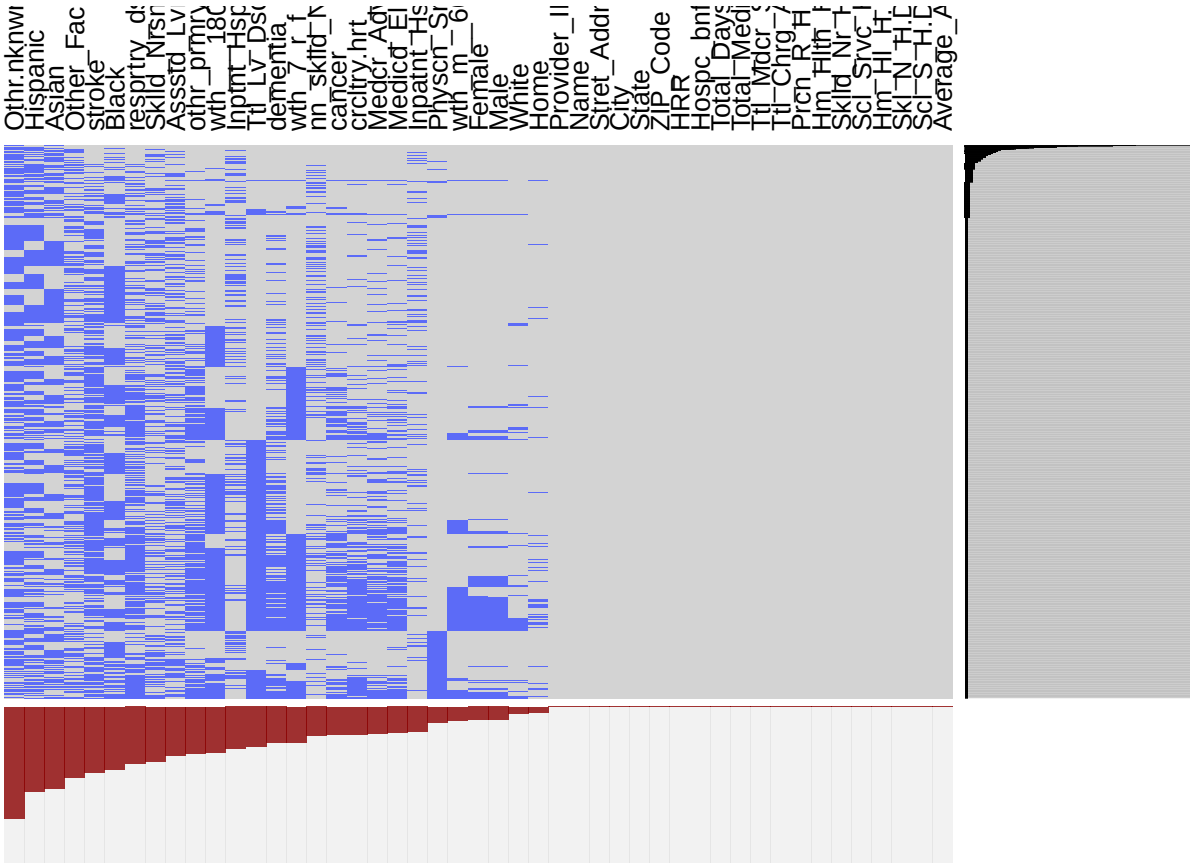
Our missing data analysis indicate that most missing variables are related to race as can be obesserved from the following plot. When filling paperwork to be enrolled in Hospice care , patients tend to not answer about their race and admission staff simply write down unknown and hence lots of missing in race. We also tried to identify any trends in missing data using scattermatrix missing from VIM package. The red cross marks corresponds to missing data while the blue dots depits available data. We could not find any relationships among missig data especially we further do our analysis on primary diagnosis, and site of sesrvices. Since there are more than 50% o missing data on race we did not continue our analysis on race data.

```
patientcare.data.2016C <- read.csv("./Healthcare_Provider_2016C.csv",
                                   stringsAsFactors = FALSE,
                                   na.strings = NA )
patientcare.data.2016C <- (patientcare.data.2016C[, -1])
```

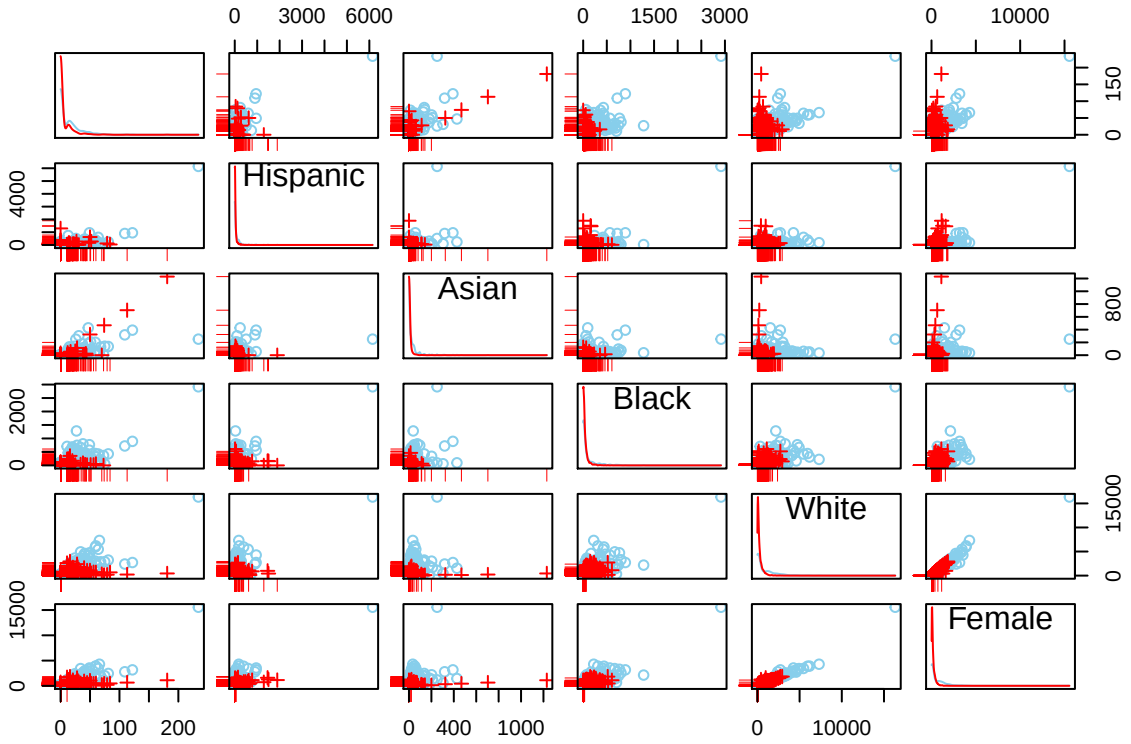
```r
patientcare.data.2016C[, 1] <- as.character(patientcare.data.2016C[, 1])
patientcare.data.2016C[c(6,8:47)] <- lapply(patientcare.data.2016C[c(6,8:47)],
                                            as.numeric)
patientcare.data.2016C <- as.data.table(patientcare.data.2016C)

na_percent <- tibble(variable = colnames(patientcare.data.2016C),
percent = 100* colSums(is.na(patientcare.data.2016C))/nrow(patientcare.data.2016C))

ggplot(na_percent,aes(fct_reorder(variable, percent), percent)) +
  geom_col(color = "grey50", fill = "lightblue") + coord_flip() +
  ggtitle("Percent missing by variable") + labs(x="Variables",
                                                y = "Percentage") + theme_grey(9)
```



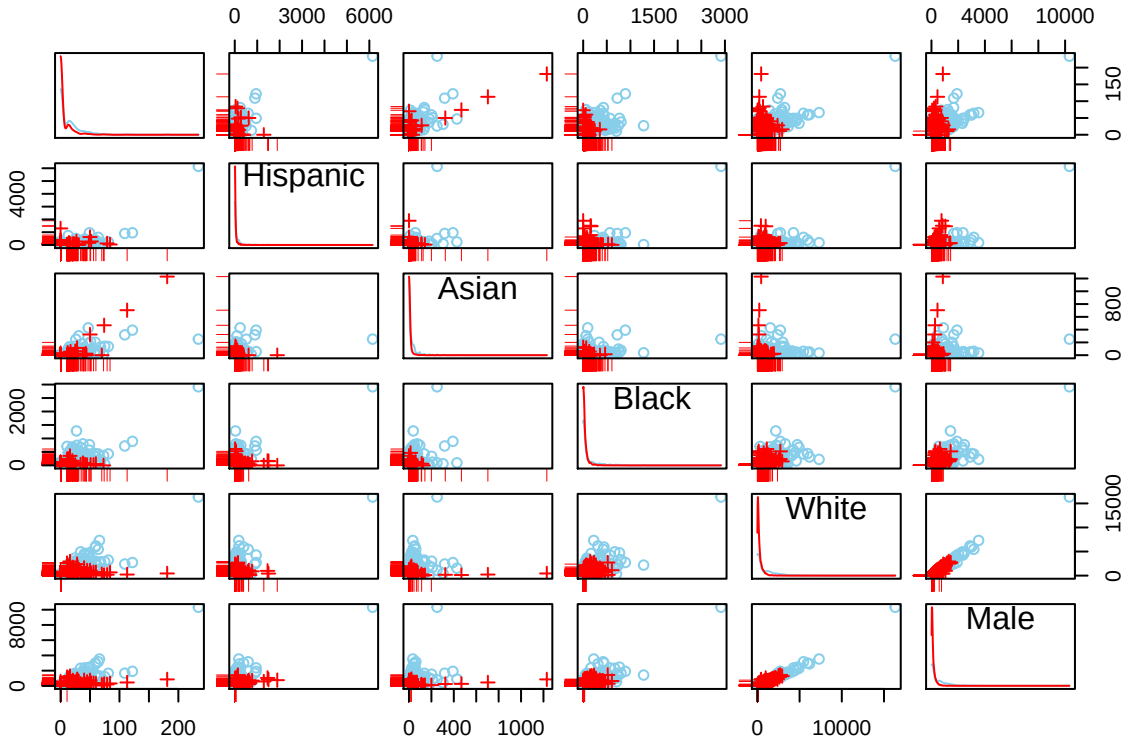

```r
extracat::visna (patientcare.data.2016C, sort = "b")
```



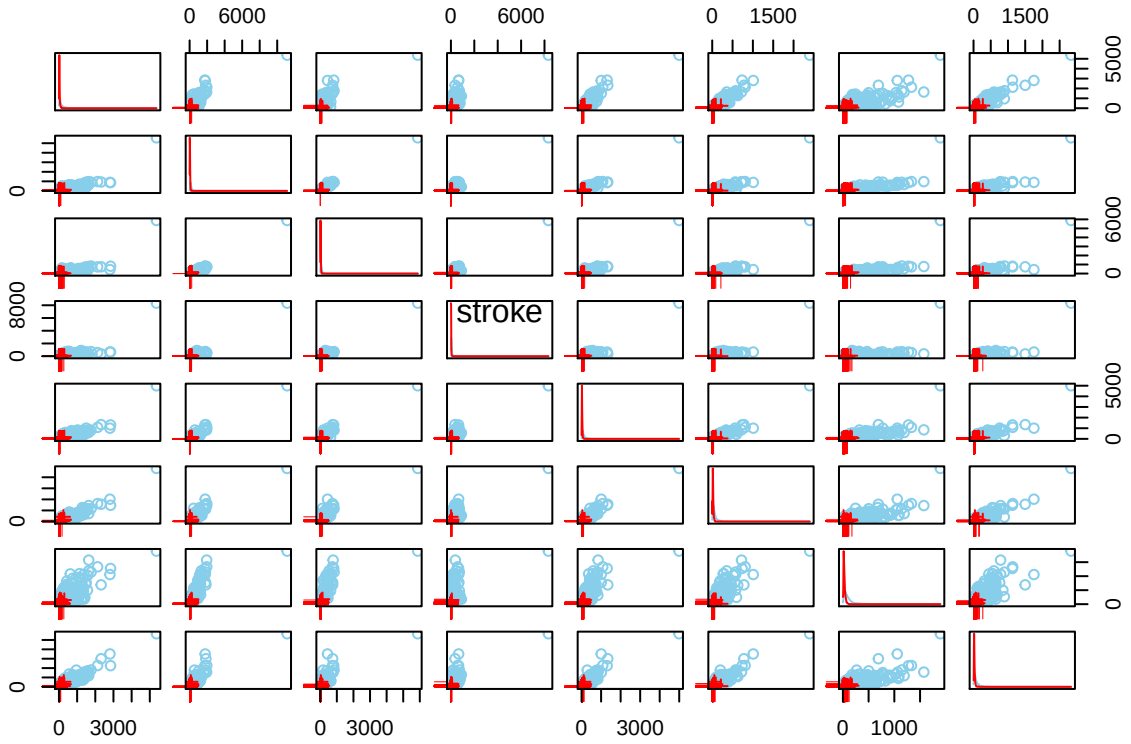

```
patientcare.data.2016C <- as.data.frame(patientcare.data.2016C)

scattmatrixMiss(patientcare.data.2016C[, c("Other.unknown_race", "Hispanic", "Asian", "Black", "White",
```




```
scattmatrixMiss(patientcare.data.2016C[, c("Other.unknown_race", "Hispanic", "Asian", "Black", "White",
```




```
scattmatrixMiss(patientcare.data.2016C[, c("with_7_or_fewer_CDys", "with_more_than_60_CDys", "with_more_
```



4. Main analysis (Exploratory Data Analysis)

Our analysis is five fold. In the following sections we focus our analysis systematically on (1)Medicare Payments, (2)Length of stay, (3)Site of Service and Care Visits, (4)Trends in primary diagnosis and site of service, and (5)possible Fraud prediction respectively.

```
## 2014
total.hospice.patient.2014 <- hospice.data.2014[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = "Hospi
total.hospice.patient.state.2014 <- hospice.data.2014[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols =
print(sprintf("Total Number of patients in the Hospice in year 2014 %d",total.hospice.patient.2014[,Hos
```

```
## [1] "Total Number of patients in the Hospice in year 2014 1372087"
```

```
total.patients.2014 <- total.hospice.patient.2014[,Hospice_beneficiaries]
## 2015
total.hospice.patient.2015 <- hospice.data.2015[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = "Hospi
#total.hospice.patient.2016.remove.566 <- hospice.data.2016.remove.566[,lapply(.SD, FUN = sum, na.rm =
total.hospice.patient.state.2015 <- hospice.data.2015[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols =
print(sprintf("Total Number of patients in the Hospice in year 2015 %d",total.hospice.patient.2015[,Hos
```

```
## [1] "Total Number of patients in the Hospice in year 2015 1431448"
```

```
total.patients.2015 <- total.hospice.patient.2015[,Hospice_beneficiaries]
total.hospice.patient.2016 <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = "Hospi
total.hospice.patient.state.2016 <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols =
print(sprintf("Total Number of patients in the Hospice in year 2016 %d",total.hospice.patient.2016[,Hos
```

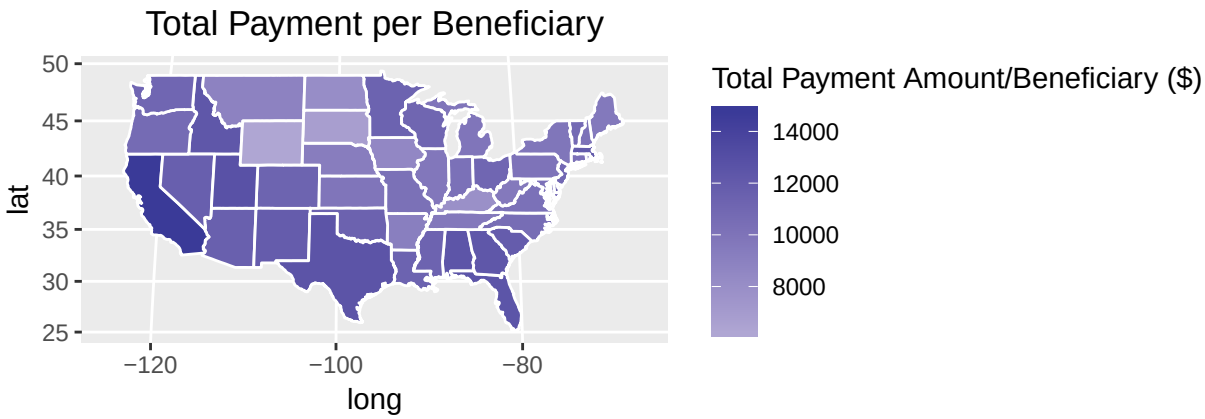
```
## [1] "Total Number of patients in the Hospice in year 2016 1476477"
```

```
total.patients.2016 <- total.hospice.patient.2016[,Hospice_beneficiaries]
```

4.1 Medicare Payments

When we first visualized the hospice beneficiaries for fiscal year 2016 we observed certain states, namely California, Texas, and Florida has the highest Medicare paymentsbut when we compare with the U.S. population (https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population), that is not a surprise because those states have high population.

```
library("mapproj")
library(maps)


state_borders <- map_data("state")
state_name <- data.frame(state.name,state.abb)

total.payment.state.2016 <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = c("Total

total.payment.state.2016.per.person <- mutate(total.payment.state.2016, Payment.per.beneficiaries = Tota


payment_plot <-total.payment.state.2016.per.person %>% left_join(state_name, by = c("State" = "state.ab

payment_plot$state.name = tolower(payment_plot$state.name)
payment_plot = as.data.table(payment_plot)

state_borders_payment<- state_borders %>% left_join(payment_plot, by = c("region" = "state.name"))

ggplot(state_borders_payment) +
                            geom_polygon(aes(x = long, y = lat, group = group,
                            fill = `Payment.per.beneficiaries`), color = "white") +
                            scale_fill_gradient2() +
                            coord_map("gilbert") +
                            labs(fill = "Total Payment Amount/Beneficiary ($) ",
                            title = "Total Payment per Beneficiary") +
                            theme(panel.border =  element_blank()) + theme(plot.title = element_text(hju
```



From the US map of Medicare payment per beneficiary by state, we can see that above mentioned states for example, Florida does not stand out any more as in Medicare payment per provider, which suggests that the reason Florida gets high payment amount is because a large population of people are using the Medicare service in Florida, instead of the high amount of payment for each beneficiary. California has the highest Medicare payment per beneficiary while North Dakota has the lowest. The top three cities with highest Medicare payment per beneficiary are all in California.

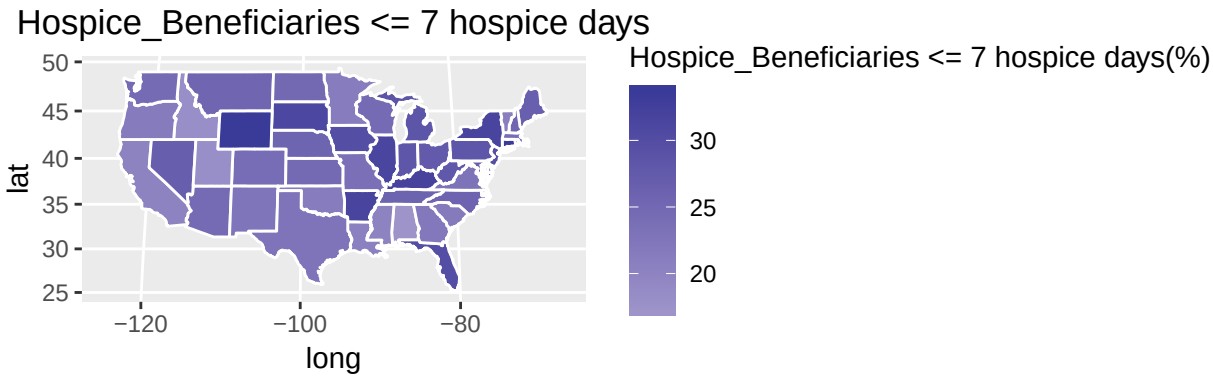
4.2 Length of stay

It is interesting to explore the distribution of Medicare beneficiaries' hospice stay which is short stay (less than 7 days) and long stay (greater than 180 days). Hospice length of stay will have an impact for the Medicare cost and for the quality of care. For example, according to the below figures New York beneficiaries spend less than 7 days while Alabama beneficiaries spending more than 180 days in hospice care.

4.2.1 Length of stay less than 7 days

```
state_borders <- map_data("state")
state_name <- data.frame(state.name,state.abb)

care.days.variables.7day <- c("Hospice_beneficiaries_with_7_or_fewer_hospice_care_days")
care.days.2016.7day <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = care.days.vari
care.days.2016.num <- hospice.data.2016[,.N, by = State]
```

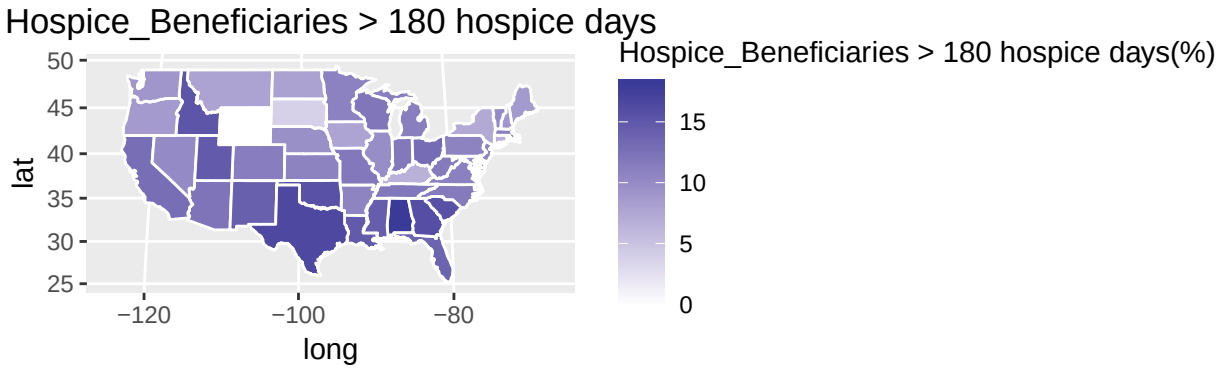
```r
care.days.2016.7day  <- care.days.2016.7day[care.days.2016.num , on = "State"]



care.days.2016.7day <- care.days.2016.7day[total.hospice.patient.state.2016, on = "State"]
care.days.2016.7day[, `:=`(Percentage.7day,Hospice_beneficiaries_with_7_or_fewer_hospice_care_days*100/


# plot < 7 days care

care_days_plot_7day <- care.days.2016.7day[,.(State,N,Percentage.7day )]
care_days_plot_7day<- care_days_plot_7day[,.(State,Percentage.7day)]

care_days_plot_7day <-care_days_plot_7day %>%  left_join(state_name, by = c("State" = "state.abb"))

care_days_plot_7day$state.name = tolower(care_days_plot_7day$state.name)
care_days_plot_7day = as.data.table(care_days_plot_7day)


state_borders_care_days_7day <- state_borders %>% left_join(care_days_plot_7day , by = c("region" = "st

ggplot(state_borders_care_days_7day) +
                          geom_polygon(aes(x = long, y = lat, group = group,
                          fill = `Percentage.7day`), color = "white") +
                          scale_fill_gradient2() +
                          coord_map("gilbert") +
                          labs(fill = "Hospice_Beneficiaries <= 7 hospice days(%)", title = "Hospice_Ben
                          theme(panel.border =  element_blank()) + theme(plot.title = element_text(hjust
```



4.2.2 Length of stay more than 180 days

```
library("mapproj")
library(maps)


state_borders <- map_data("state")
state_name <- data.frame(state.name,state.abb)

care.days.variables.180day <- c("Hospice_beneficiaries_with_more_than_180_hospice_care_days")
care.days.2016.180day <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = care.days.va

care.days.2016.num <- hospice.data.2016[,.N, by = State]

care.days.2016.180day  <- care.days.2016.180day[care.days.2016.num , on = "State"]

care.days.2016.180day <- care.days.2016.180day[total.hospice.patient.state.2016, on = "State"]

care.days.2016.180day[, `:=`(Percentage.180day,Hospice_beneficiaries_with_more_than_180_hospice_care_da


# plot >180 days care
```

```
care_days_plot_180day <- care.days.2016.180day[,.(State,N,Percentage.180day )]
care_days_plot_180day<- care_days_plot_180day[,.(State,Percentage.180day)]


care_days_plot_180day <-care_days_plot_180day %>% left_join(state_name, by = c("State" = "state.abb"))

care_days_plot_180day$state.name = tolower(care_days_plot_180day$state.name)
care_days_plot_180day = as.data.table(care_days_plot_180day)


state_borders_care_days_180day <- state_borders %>%  left_join(care_days_plot_180day , by = c("region" =

ggplot(state_borders_care_days_180day) +
                     geom_polygon(aes(x = long, y = lat, group = group,
                    fill = `Percentage.180day`), color = "white") + scale_fill_gradient2() +
                    coord_map("gilbert") +
                    labs(fill = "Hospice_Beneficiaries > 180 hospice days(%)",  title = "Hospice_Benefici
                    theme(panel.border =  element_blank()) + theme(plot.title = element_text(hjust = 0.5)
```

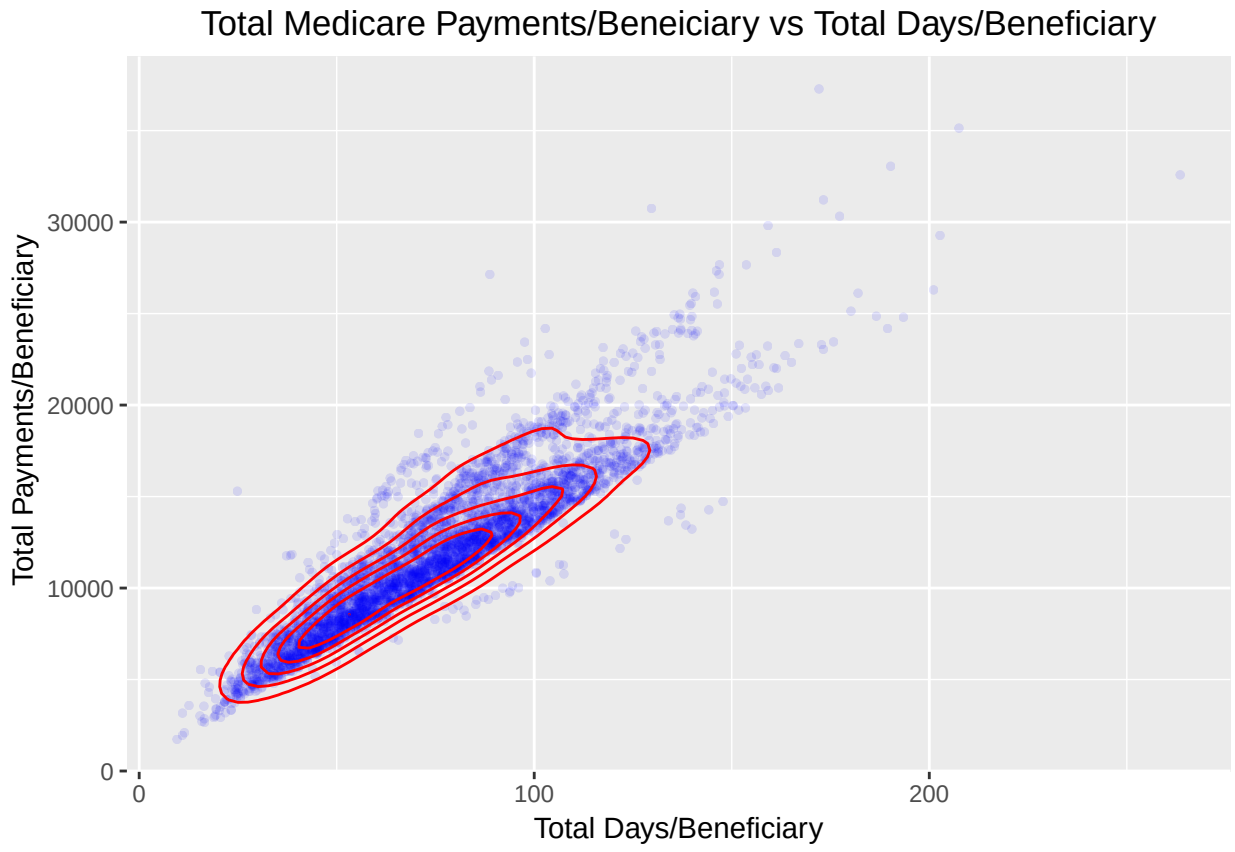


Hospice care was designed to provide a better quality of life to the patients and have less burden on the Medicare system. Hospice care is given to patients who are terminally ill with a life expectancy of 6 months or less and certified by two medical practitioners (https://www.medicare.gov/coverage/hospice-and-respite-care.html).Thus Hospice benficiaries staying long than 180 days (6 months) indeed a redflag of possible misuse of public benefits and gives us an area to explore more. On the other hand, a hospice beneficiary is considered to have a live discharge if hospice beneficiary did not die during hospice care. Thus, live discharge can be identified as a potential weakness in the Medicare hospice benefit. Latter part of the analysis we dedicate to this findings.

4.2.3 Medicare Payents and Length of Stay

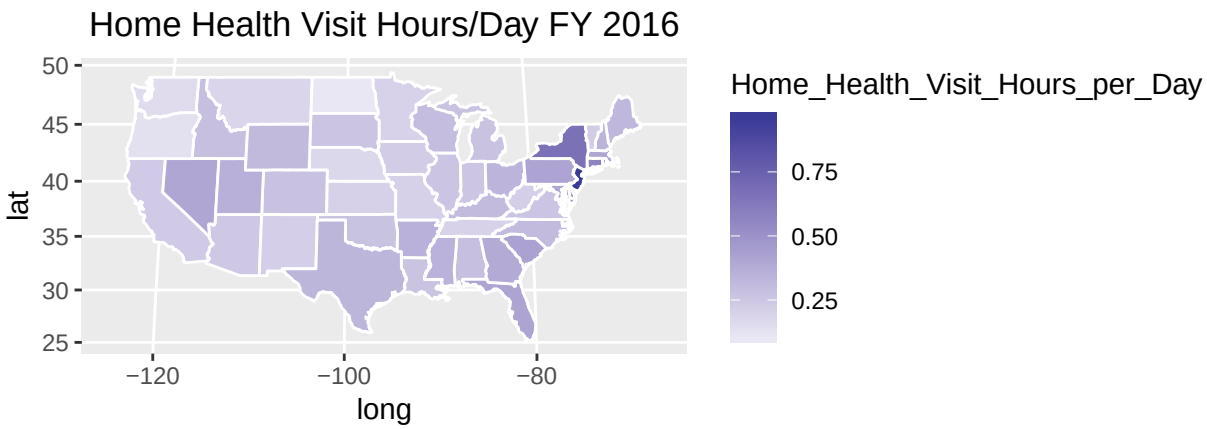
The straight diagonal lines in the below plot and associated conutours indicate a linear relation between total medicare payment amount total days in care facilities.

```
#hospice.data.2016.edit <- hospice.data.2016[-994,]

hospice.data.2016.edit <- hospice.data.2016

ggplot(hospice.data.2016.edit, aes(x=Total_Days/Hospice_beneficiaries , y=Total_Medicare_Payment_Amount,
geom_point(color = "blue", alpha = 0.1, stroke = 0) + labs( title = "Total Medicare Payments/Beneiciary
```

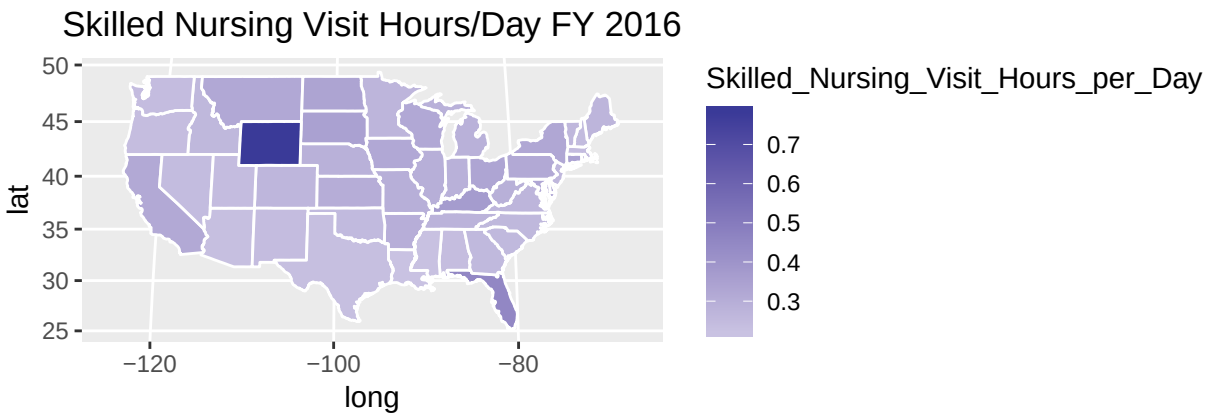


4.3 Site of Service and Care Visits

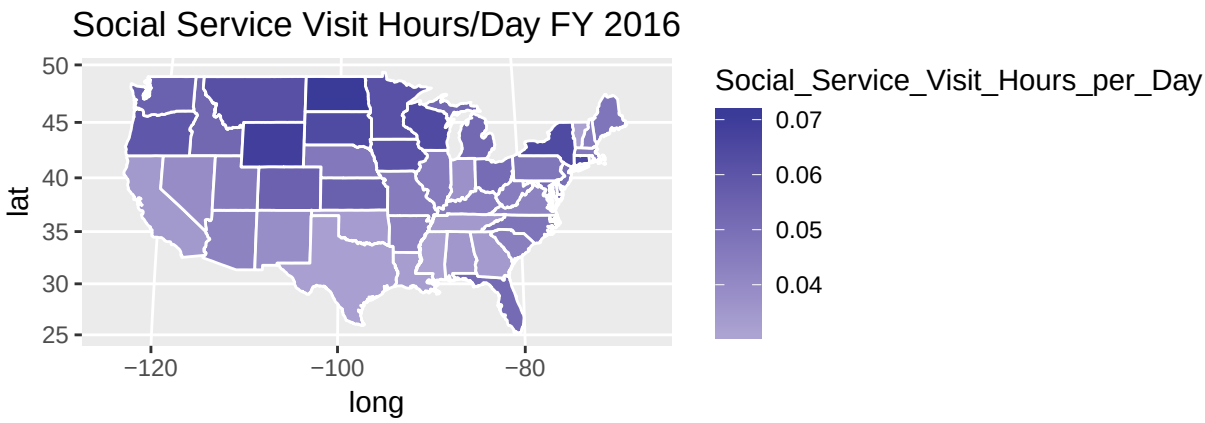
```
library("mapproj")
library(maps)


state_borders <- map_data("state")
state_name <- data.frame(state.name,state.abb)


care.type.variables <- c("Home_Health_Visit_Hours_per_Day","Skilled_Nursing_Visit_Hours_per_Day","Social
care.type.2016 <- hospice.data.2016[,lapply(.SD, FUN = sum, na.rm = TRUE), .SDcols = care.type.variables

care.type.2016.num <- hospice.data.2016[,.N, by = State]

care.type.2016 <- care.type.2016[care.type.2016.num, on = "State"]

# plot home health

care_type_plot_hh <- care.type.2016[,.(State,N,Home_Health_Visit_Hours_per_Day )]

care_type_plot_hh <- care_type_plot_hh [,`:=`(Home_Health_Visit_Hours_per_Day,Home_Health_Visit_Hours_pe

care_type_plot_hh <-care_type_plot_hh[,.(State,Home_Health_Visit_Hours_per_Day)]

care_type_plot_hh <-care_type_plot_hh %>% left_join(state_name, by = c("State" = "state.abb"))

care_type_plot_hh$state.name = tolower(care_type_plot_hh$state.name)
care_type_plot_hh= as.data.table(care_type_plot_hh)


state_borders_care_type_hh<- state_borders %>% left_join(care_type_plot_hh, by = c("region" = "state.nam

ggplot(state_borders_care_type_hh) +
                            geom_polygon(aes(x = long, y = lat, group = group, fill = `Home_Health_Visit
                            scale_fill_gradient2() +
                            coord_map("gilbert") +
                            labs(fill = "Home_Health_Visit_Hours_per_Day ",
                            title = "Home Health Visit Hours/Day FY 2016") +
                            theme(panel.border =  element_blank()) + theme(plot.title = element_text(hju
```




```
# plot Skill nurse

care_type_plot_sn <- care.type.2016[,.(State,N,Skilled_Nursing_Visit_Hours_per_Day )]

care_type_plot_sn <- care_type_plot_sn[,`:=`(Skilled_Nursing_Visit_Hours_per_Day,Skilled_Nursing_Visit_

care_type_plot_sn <-care_type_plot_sn[,.(State,Skilled_Nursing_Visit_Hours_per_Day)]


care_type_plot_sn <-care_type_plot_sn %>% left_join(state_name, by = c("State" = "state.abb"))

care_type_plot_sn$state.name = tolower(care_type_plot_sn$state.name)
care_type_plot_sn= as.data.table(care_type_plot_sn)


state_borders_care_type_sn<- state_borders %>% left_join(care_type_plot_sn, by = c("region" = "state.na

ggplot(state_borders_care_type_sn) +
                                  geom_polygon(aes(x = long, y = lat, group = group,
                                  fill = `Skilled_Nursing_Visit_Hours_per_Day`), color = "white") +
                                  scale_fill_gradient2() +
                                  coord_map("gilbert") +
                                  labs(fill = "Skilled_Nursing_Visit_Hours_per_Day ",  title = "Skilled
                                    theme(panel.border =  element_blank()) + theme(plot.title = element_
```




```
# plot social services

care_type_plot_ss <- care.type.2016[,.(State,N,Social_Service_Visit_Hours_per_Day )]

care_type_plot_ss <- care_type_plot_ss[,`:=`(Social_Service_Visit_Hours_per_Day,Social_Service_Visit_Hou

care_type_plot_ss <-care_type_plot_ss[,.(State,Social_Service_Visit_Hours_per_Day)]


care_type_plot_ss <-care_type_plot_ss %>% left_join(state_name, by = c("State" = "state.abb"))

care_type_plot_ss$state.name = tolower(care_type_plot_ss$state.name)
care_type_plot_ss= as.data.table(care_type_plot_ss)


state_borders_care_type_ss<- state_borders %>% left_join(care_type_plot_ss, by = c("region" = "state.nam

ggplot(state_borders_care_type_ss) +
                                geom_polygon(aes(x = long, y = lat, group = group,
                                fill = `Social_Service_Visit_Hours_per_Day`), color = "white") +
                                scale_fill_gradient2() + coord_map("gilbert") +
                                labs(fill = "Social_Service_Visit_Hours_per_Day ",title = "Social Servic
                                theme(panel.border =  element_blank()) + theme(plot.title = element_text
```



##4.4 Trends in primary diagnosis and site of service

4.4.1 Hospice beneficiaries' primary diagnosis

When we look the distribution of hospice beneficiaries based on primary diagnosis, cancer is the foremost primary diagnosis for the hospice patients in all three years as shown in the following Alluvial flow diagram. Other leading diagnosis are Dementia and Heart Disease.

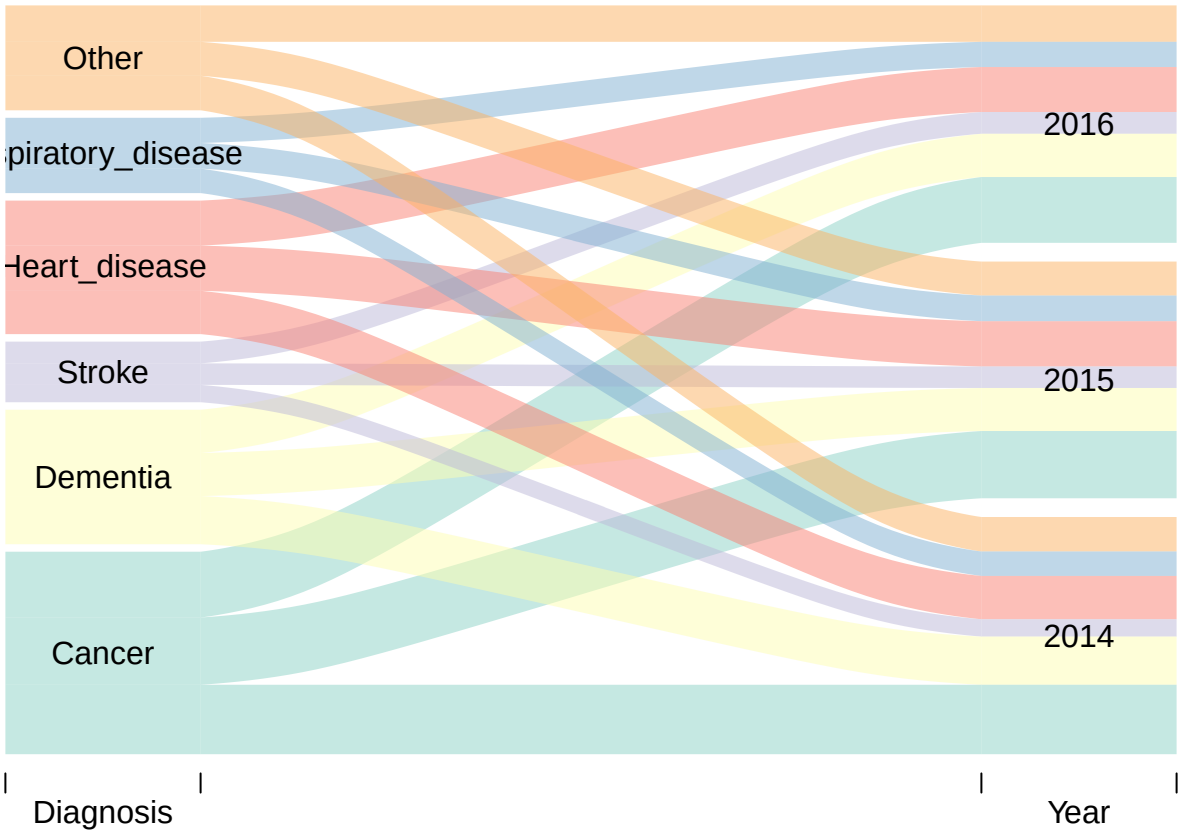
```
# 2016

select.2016 <- hospice.data.2016[,c(35:47)]


select.variables.2016 <- colnames(hospice.data.2016[,35:47])
new.variable.2016 <- c("Cancer", "Dementia", "Stroke","Heart_disease","respiratory_disease","Other","Ho
select.factors.2016 <- select.2016[, .SDcols = select.variables.2016]


sum.select.2016 <- select.factors.2016 [, lapply(X = .SD, FUN = sum, na.rm = TRUE), .SDcols = select.va

y <- hospice.data.2016[, lapply(X = .SD, FUN = sum, na.rm = TRUE), .SDcols = select.variables.2016 , by

percentage.select.2016 <- round(sum.select.2016[1,]/total.hospice.patient.2016[,Hospice_beneficiaries]*

setnames(percentage.select.2016,  select.variables.2016, new.variable.2016)
```

```
diagnosis.2016 <- percentage.select.2016[,c(1:6)]
site.2016 <- percentage.select.2016[,c(7:13)]


diagnosis.2016[, `:=`(Metric, c("Per"))]
diagnosis.2016 <- dcast(melt( diagnosis.2016, id.vars = "Metric"), variable ~ Metric)
setnames(diagnosis.2016, old = c("variable","Per"), new = c("Diagnosis","Per"))

site.2016[, `:=`(Metric, c("Per"))]
site.2016 <- dcast(melt( site.2016, id.vars = "Metric"), variable ~ Metric)
setnames(site.2016, old = c("variable","Per"), new = c("Site_of_Service","Per"))

diagnosis.2016 <- as.data.table(diagnosis.2016)
diagnosis.2016 <-  diagnosis.2016[, Year:= c(2016,2016,2016,2016,2016,2016)]

site.2016 <- as.data.table(site.2016)
site.2016 <-  site.2016[, Year:= c(2016,2016,2016,2016,2016,2016,2016)]

#2015

select.2015 <- hospice.data.2015[,c(35:47)]


select.variables.2015 <- colnames(hospice.data.2015[,35:47])
new.variable.2015 <- c("Cancer", "Dementia", "Stroke","Heart_disease","respiratory_disease","Other","Ho
select.factors.2015 <- select.2015[, .SDcols = select.variables.2015]


sum.select.2015 <- select.factors.2015 [, lapply(X = .SD, FUN = sum, na.rm = TRUE), .SDcols = select.va
percentage.select.2015 <- round(sum.select.2015[1,]/total.hospice.patient.2015[,Hospice_beneficiaries]*

setnames(percentage.select.2015,  select.variables.2015, new.variable.2015)

diagnosis.2015 <- percentage.select.2015[,c(1:6)]
site.2015 <- percentage.select.2015[,c(7:13)]


diagnosis.2015[, `:=`(Metric, c("Per"))]
diagnosis.2015 <- dcast(melt( diagnosis.2015, id.vars = "Metric"), variable ~ Metric)
setnames(diagnosis.2015, old = c("variable","Per"), new = c("Diagnosis","Per"))

site.2015[, `:=`(Metric, c("Per"))]
site.2015 <- dcast(melt( site.2015, id.vars = "Metric"), variable ~ Metric)
setnames(site.2015, old = c("variable","Per"), new = c("Site_of_Service","Per"))

diagnosis.2015 <- as.data.table(diagnosis.2015)
diagnosis.2015 <-  diagnosis.2015[, Year:= c(2015,2015,2015,2015,2015,2015)]

site.2015 <- as.data.table(site.2015)
site.2015 <-  site.2015[, Year:= c(2015,2015,2015,2015,2015,2015,2015)]

#2014
```

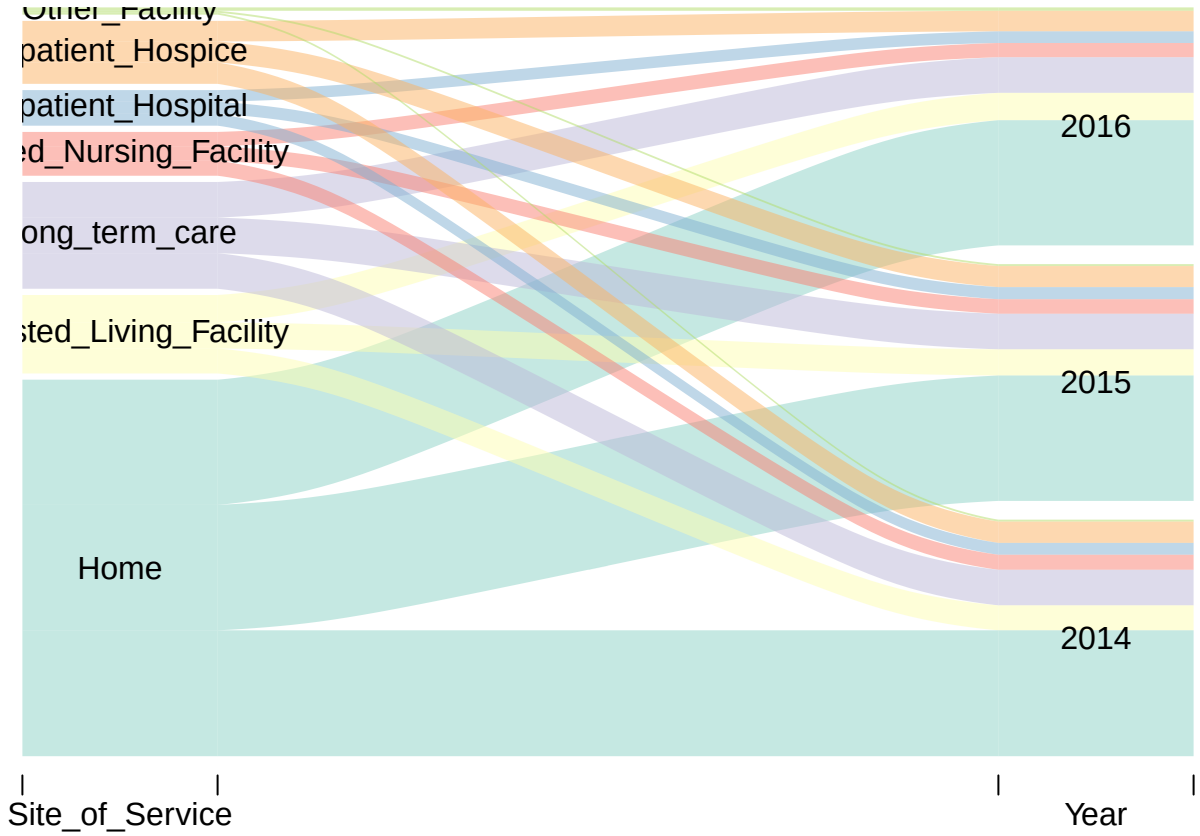
```
select.2014 <- hospice.data.2014[,c(35:47)]


select.variables.2014 <- colnames(hospice.data.2014[,35:47])
new.variable.2014 <- c("Cancer", "Dementia", "Stroke","Heart_disease","respiratory_disease","Other","Hos
select.factors.2014 <- select.2014[, .SDcols = select.variables.2014]


sum.select.2014 <- select.factors.2014[, lapply(X = .SD, FUN = sum, na.rm = TRUE), .SDcols = select.vari
percentage.select.2014 <- round(sum.select.2014[1,]/total.hospice.patient.2014[,Hospice_beneficiaries]*

setnames(percentage.select.2014,  select.variables.2014, new.variable.2014)

diagnosis.2014 <- percentage.select.2014[,c(1:6)]
site.2014 <- percentage.select.2014[,c(7:13)]


diagnosis.2014[, `:=`(Metric, c("Per"))]
diagnosis.2014 <- dcast(melt( diagnosis.2014, id.vars = "Metric"), variable ~ Metric)
setnames(diagnosis.2014, old = c("variable","Per"), new = c("Diagnosis","Per"))

site.2014[, `:=`(Metric, c("Per"))]
site.2014 <- dcast(melt( site.2014, id.vars = "Metric"), variable ~ Metric)
setnames(site.2014, old = c("variable","Per"), new = c("Site_of_Service","Per"))

diagnosis.2014 <- as.data.table(diagnosis.2014)
diagnosis.2014 <-  diagnosis.2014[, Year:= c(2014,2014,2014,2014,2014,2014)]

site.2014 <- as.data.table(site.2014)
site.2014 <-  site.2014[, Year:= c(2014,2014,2014,2014,2014,2014,2014)]
```

```
data.dignose <- bind_rows(diagnosis.2014,diagnosis.2015,diagnosis.2016)
data.dignose <- data.dignose[,c(3,1,2)]

data.site <- bind_rows(site.2014,site.2015,site.2016)
data.site <- data.site[,c(3,1,2)]
```

```
library(alluvial)
pal <- RColorBrewer::brewer.pal(15, "Set3")

alluvial(data.dignose[, 2:1], freq = data.dignose$Per,
         blocks = FALSE,
         alpha = 0.5,
         col = pal[match(data.dignose$Diagnosis,
                         unique(data.dignose$Diagnosis)) ])
```




```
library(alluvial)
pal <- RColorBrewer::brewer.pal(10, "Set3")

alluvial(data.site[, 2:1], freq = data.site$Per,
         blocks = FALSE,
         alpha = 0.5,
         col = pal[match(data.site$Site_of_Service,
                         unique(data.site$Site_of_Service)) ])
```



4.4.2 Hospice beneficiaries Site of Service Trend

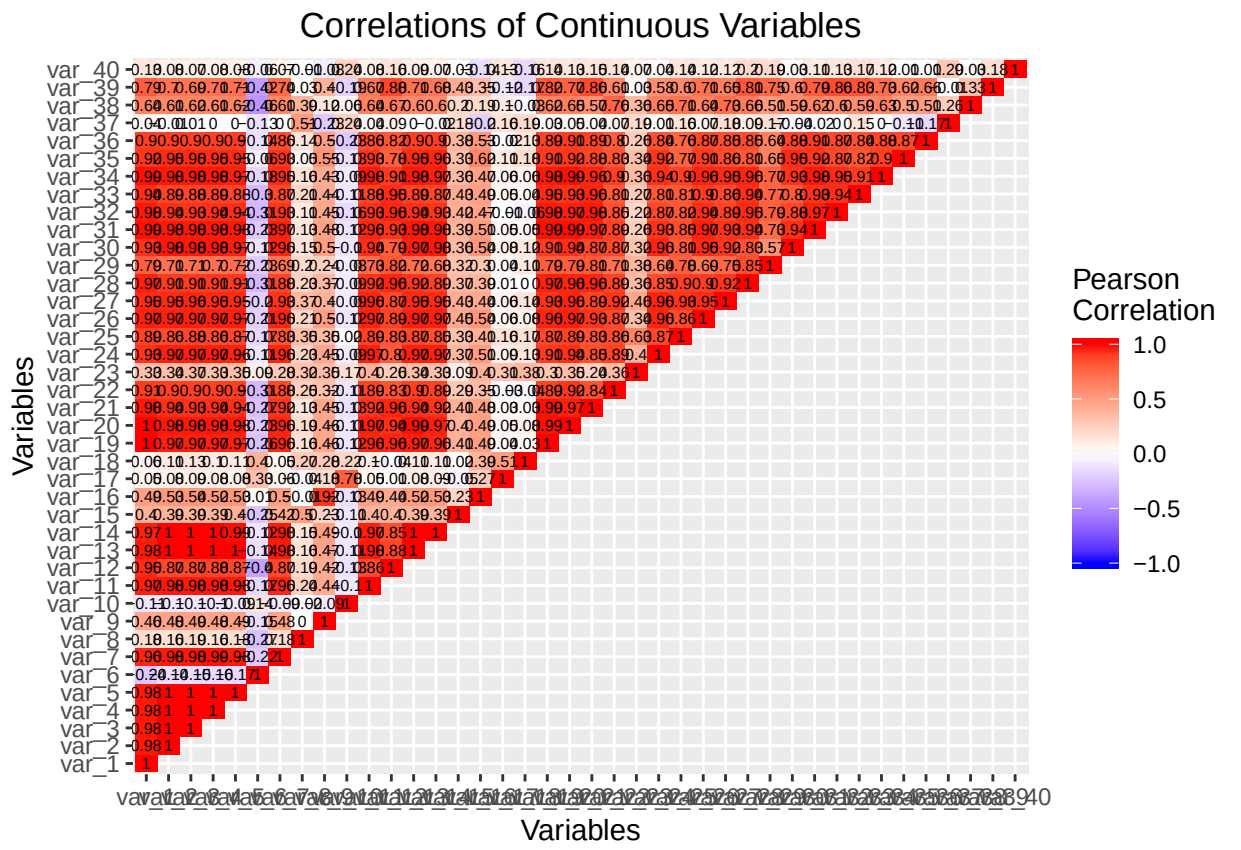
We analyzed the number of hospice beneficiaries based on the site of service to identify their distribution for the 3 years. Alluvial flow diagram depicts the allocation of services in different sites. Based on the analysis it is evident that the desired site of service for the hospice beneficiaries is cared at home followed by long-term care/non-skilled nursing facilities and assisted living facilities.

4.5 Fraud prediction

The graph shows pairwise correlation between continuous variables in our dataset. The correlation is between -1 and 1. Highly positive correlation is denoted by blue while highly negative correlation is denoted by red. Most pair of variables exhibit highly positive correlation, which could be a potential problem for a linear regression model, since linear regression assumes no perfect collinearity. Also, linear regression behaves poorly in high dimensions. So, we think lasso regression is a good choice for a possible model.Similar to linear regression, lasso regression finds the coefficients using least squared error. The only difference is that lasso regression adds a penalty term to avoid overfitting.

```
dat.respondent<- hospice.data.2016[,c(8:47)]
chart = as.data.table(data.frame(variable_name = colnames(dat.respondent), abbr = paste("var",as.charac
setnames(x = dat.respondent, old = colnames(dat.respondent), new = paste("var",as.character(1:length(co

# Get upper triangle of the correlation matrix
```

```r
get_upper_tri <- function(cormat){
  cormat[lower.tri(cormat)]<- NA
  return(cormat)
}

correlation_matrix <- cor(dat.respondent, use = "complete.obs")
upper_tri <- get_upper_tri(correlation_matrix)
melted_cormat <- melt(upper_tri, na.rm = TRUE)
melted_cormat$value <- round.numerics(melted_cormat$value,2)

ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
              geom_tile() +
              geom_text(aes(label = value),size = 2) +
              scale_fill_gradient2(low = "blue", high = "red", mid = "white",
              midpoint = 0, limit = c(-1,1), name="Pearson\nCorrelation") +
              labs(x = "Variables", y = "Variables",
              title = "Correlations of Continuous Variables",
              fill = "Correlation") +
              theme(plot.title = element_text(hjust = 0.5))
```



One of the important applications of the lasso regression is to model on total Medicare payment and to detect whether a claim is a fraud or not. We leave the deatiled analysis to a future project. But the idea is nortworthy and the explained here. For example, for a new observation, we can apply our fitted model to predict total payment amount, and set the threshold to be:

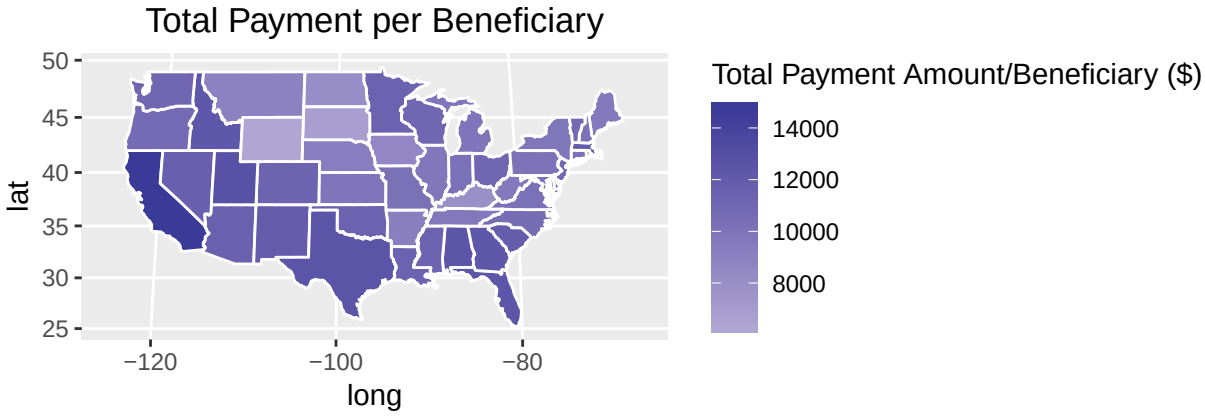
"Threshold = n x Standard Deviation of Residuals + Predicted Payment"

(Here selection of n can be detremined by exact threshold one would like to have). If the actual payment amount exceeds the threshold, we will flag the claim to be a fraud.

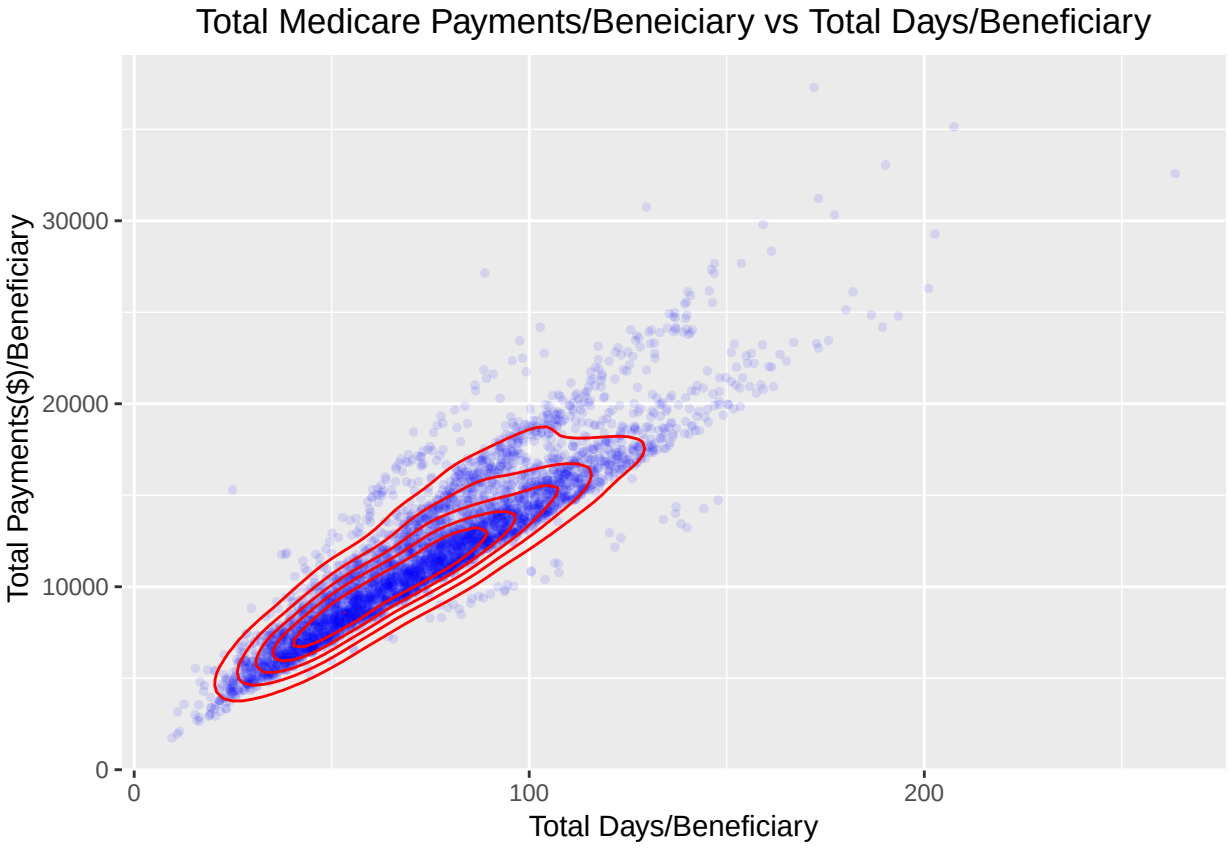
5. Executive summary

Hospice care was designed to provide a better quality of life to the patients and have less burden on the Medicare system. Hospice care is given to patients who are terminally ill with a life expectancy of 6 months or less and certified by two medical practitioners.

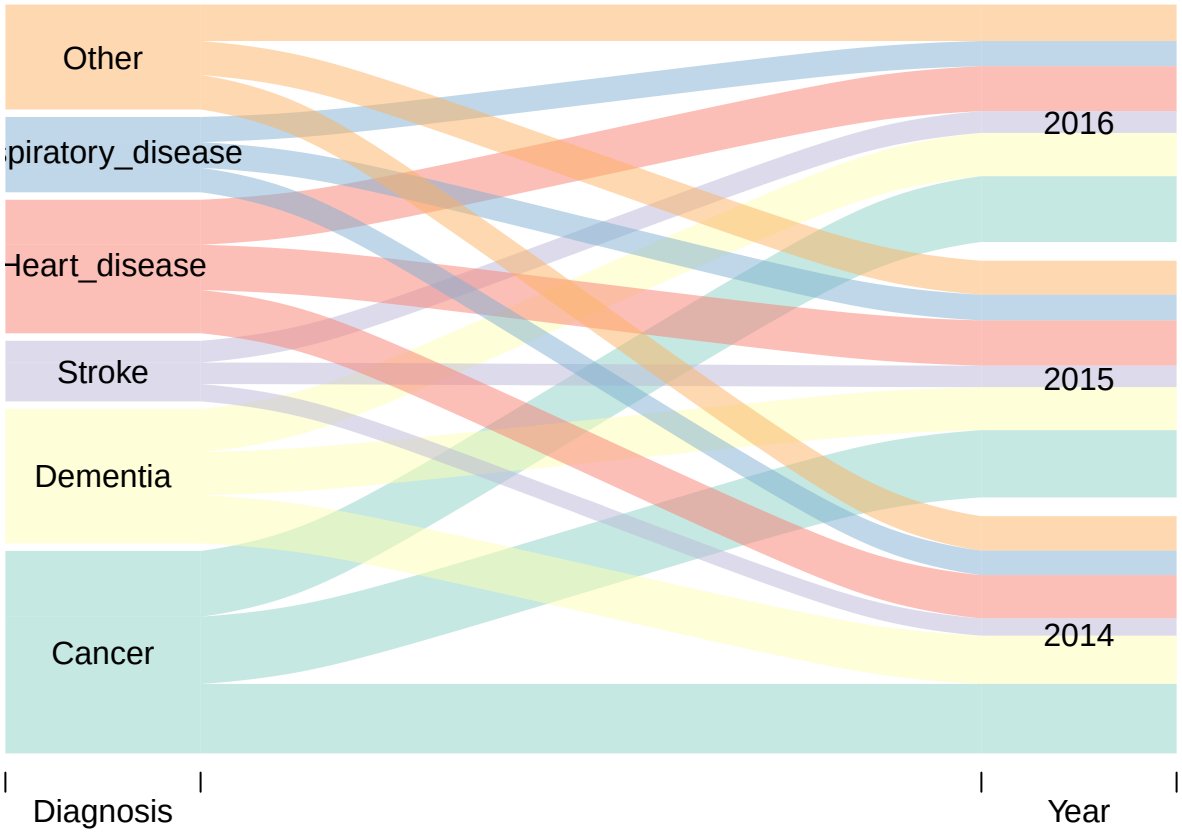
With the available hospice care data for the years 2014, 2015, and 2016, we present trends in hospice care and identify hospice care and service utilization.



From the US map of Medicare payment per beneficiary by state we observe unbiased medicare payment trends that are not effect by respective populations in highly populated areas. California has the highest Medicare payment per beneficiary while North Dakota has the lowest.



The straight diagonal lines in the above scatter plot and associated conutours indicate a linear relation between total medicare payment amount total days in care facilities. This is an expected trend and it is unbiased to the total benificiary population. This is one other trend that indicates a hospice beneficiary is considered to have a live discharge if hospice beneficiary did not die during hospice care. Thus, live discharge can be identified as a potential weakness in the Medicare hospice benefit especially the total days that exceeds 180 days can be observed in the above plot.This is an indication of howproviders may try to exploit hospice care to be profitable. More fraudulently, providers may admit beneficiaries who may actually have more than 6 months of life expectancy and be profitable with less care provided.Thus, hospice is also known as "business of dying" according to the Washington Post (https://www.washingtonpost.com/sf/business/collection/business-of-dying/?noredirect=on&utm_term=.51fcabc55ead) due to the influx of for-profit hospice care providers and potential frauds.



Alluvial floor diagram, or Parallel coordinate plot, presents us the overall information we would be interested in and how the data presnted in three different years related to the main categories of Hospice beneficiaries' primary diagnosis and site of service. When we look at the distribution of hospice beneficiaries based on primary diagnosis, cancer is the foremost primary diagnosis for the hospice patients in all three years as shown in the above Alluvial flow diagram. Other leading diagnosis are Dementia and Heart Disease.We cannot see a difference in the diagnosis of admitted Hospice care recipients during all the three years 2014, 2015, and 2016 respectively.

6. Interactive component

In order to perform the web interactive part, we pre-calculate and format the data into self-defined form.Used D3 creae interactive comnents. We then used GitHub Pages as a static site to host our interactive data visualization. This allowed us to use the GitHubio to host project pages directly from a GitHub repository.Links to our interactive components:

Medicare payments per beneficiary (https://kcw2144-ps3060.github.io/EDAV/)

Hospice care length of stay (https://kcw2144-ps3060.github.io/Days/)

7. Conclusion

The entire hospice beneficiary population in the US for the year 2014 to 2016 was analyzed to find trends in hospice care, service, and cost utilization. There is an increased need for skilled care for hospice beneficiaries during length of stay less than 7 days. Our analysis on live discharges presented a possible vulnerability in the Medicare hospice benefits as there is an increase of 0.22% in live discharge rate from the year 2014 to 2015. Incorporating a cost model to live discharges can presented a possible hospice provider claim fraud. For the benefit of the entire Medicare system, a patient should be clearly identified with more scrutiny so that such possible fraud by the provider can be eliminated.

The assumptions and limitations: Hospice data is presented for only years 2014 to 2016. We were limited in knowing yearly trends other than for the given 3 years.Hospice care data for are skewed and nonlinear. However, the presented trends did not vary even if we considered a normally distributed subpopulation of the data. Hospice care data for all 3 years were aggregated by the provider and we inferred the predictions on each beneficiary which limits our ability to purely identify trends, care, service, and cost utilization by an individual beneficiary. Care services were given as visit hours per day in the hospice data. This limits further conclusions to the trends presented in care services as a number of hours visited or the days the service provided are not known. Future Investigation. In future, we can investigate nonlinear models to better present the outcome of the cost models and underline fraud detection.