Quality of End-of-Life Cancer Care
IST 687 Project

Catherine Devine, Yodit Ayalew, Lauren Foltz, Jack Harris

Background on Data Set

- ✓ Data source: Dartmouth Institute
  - ✓ Dartmouth Atlas Project (DAP) cancer events
  - ✓ https://data.world/dartmouthatlas/cancer-patients-death
- ✓ Data from 938 hospitals including location of hospital (city & state) as well as 12 attributes surrounding treatment and outcomes at end of life for elderly cancer patients. The data spans 2003-2007.
  - ✓ Number of deaths among cancer patients assigned to hospital
  - ✓ Percent of cancer patients dying in hospital
  - ✓ Percent of cancer patients admitted to hospital during the last month of life
  - ✓ Number of hospital days per cancer patient during the last month of life
  - ✓ Percent of cancer patients admitted to ICU during the last month of life
  - ✓ Number of ICU days per cancer patient during the last month of life
  - ✓ Percent of cancer patients receiving life-sustaining treatment during the last month of life
  - ✓ Percent of cancer patients receiving chemotherapy during the last two weeks of life
  - ✓ Percent of cancer patients enrolled in hospice during the last month of life
  - ✓ Number of hospice days per cancer patient during the last month of life
  - ✓ Percent of cancer patients enrolled in hospice during the last three days of life
  - ✓ Percent of cancer patients seeing ten or more physicians during the last six months of life

Goals

- ✓ Analyze data set to gain actionable insights that would be beneficial in the HealthCare industry (to Hospitals, Insurers, Doctors, Patients, and/or Family Members).
- ✓ Look at
  - ✓ Differences in outcomes across geographies
  - ✓ Cancer death among hospitals
  - ✓ Hospital admissions during the last month of life
  - ✓ Intensive care admission trends on time
  - ✓ Chemotherapy during the last two weeks of life
- ✓ Do descriptive statistics and histograms for attributes with distribution analysis of the charts
- ✓ Show direct or inverse variations between attributes and compare the result with correlation between those attributes
- ✓ Identify strong positive and weak negative correlations and try to interpret the results
- ✓ Show linear regression for all models and see the trend for strongest negative and positive correlation from the regression

Analysis Outline

- ✓ Data Cleansing
- ✓ Descriptive Statistics & Histograms
  - ✓ Mean, Median, Standard Deviation, Skewness, and Kurtosis
- ✓ Look at variables by Hospital
- ✓ Look at variables by State
- ✓ Visualize variables onto US Map
- ✓ Run a correlation on all 12 variables
- ✓ Run Linear Regression on selected variables

Data Cleansing

- ✓ Remove unwanted rows/columns
  - ✓ Remove first row of NA data
  - ✓ Removed last row of aggregate data
  - ✓ Remove 22 columns containing confidence interval data
- ✓ Column Names
  - ✓ Rename the remaining columns (giving simplified names)
- ✓ Numberize
  - ✓ Use Numberize function to remove commas, spaces, and return data as numeric

```
Numberize<-function(inputVector){
  inputVector<-gsub(",","",inputVector)
  inputVector<-gsub(" ","",inputVector)
  return(as.numeric(inputVector)) }
```

- ✓ How to deal with missing data?
  - ✓ Create subsets and utilize na.omit.

```
DAPCancerEvents5 <- DAPCancerEvents[,5]
DAPCancerEvents5 <- na.omit(DAPCancerEvents5)
```

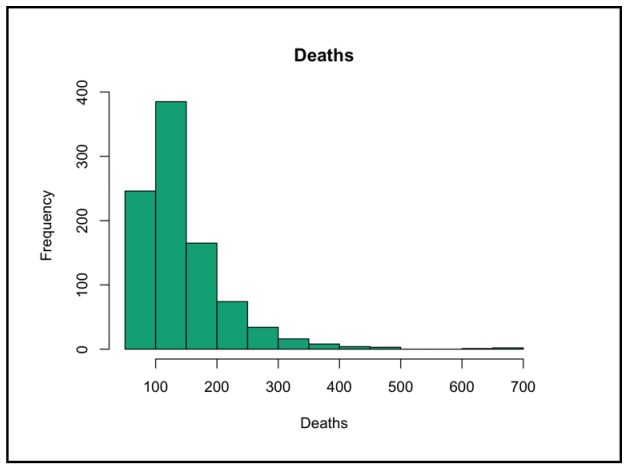
Descriptive Statistics & Histograms

- ✓ Write a function for descriptive statistics & histograms

```
descriptive <- function(x){
  print(paste("Mean:",mean(x)))
  print(paste("Median:",median(x)))
  print(paste("Standard Deviation:",sd(x)))
  print(paste("Skewness:",skewness(x)))
  print(paste("Kurtosis:",kurtosis(x)))
  hist(x,main=paste(title),xlab=title,col = cbbPalette[4])
}
```

- ✓ Apply to each of our 12 data subsets

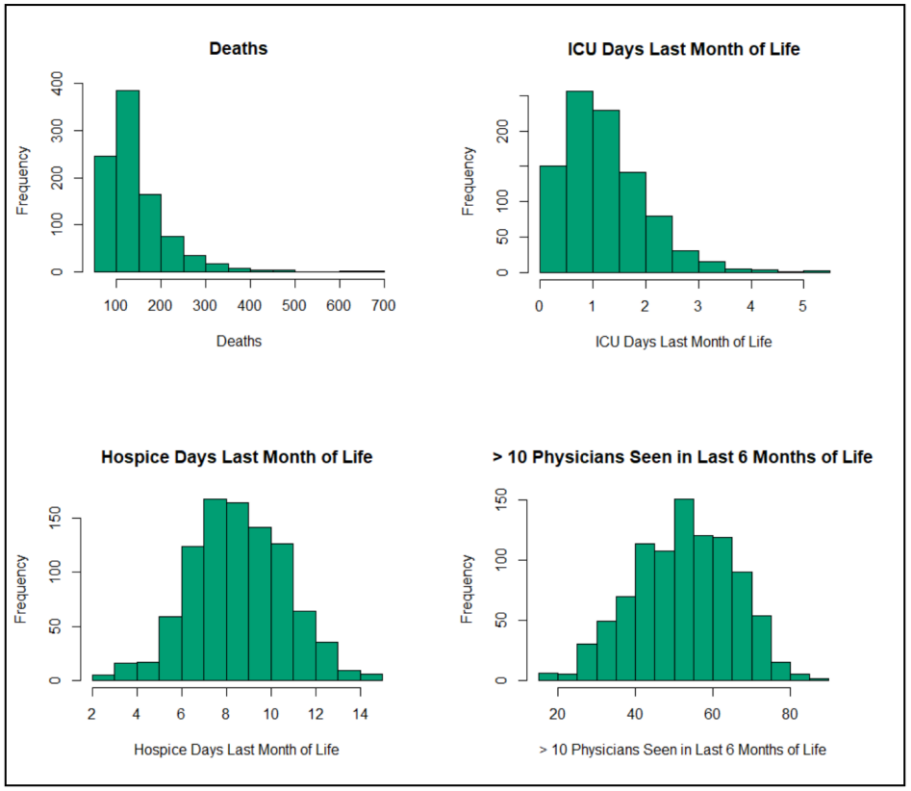
```
descriptive(DAPCancerEvents5)

## [1] "Mean: 145.265458422175"
## [1] "Median: 125"
## [1] "Standard Deviation: 69.6487827302734"
## [1] "Skewness: 2.66519410583897"
## [1] "Kurtosis: 14.3827129459792"
```



Descriptive Statistics & Histograms

Trends/Insights

- ✓ Highest Mean, Median, Standard Deviation, Skew, and Kurtosis: Deaths (note: chart shows a long right tail)
- ✓ Lowest Mean, Median, Standard Deviation: ICU days last month of Life (note: chart has a moderate right tail)
- ✓ Lowest Skew: Hospice days last month of life (note: chart shows a normal distribution)
- ✓ Lowest Kurtosis: % over 10 physicians seen in last 6 months (note: chart shows a normal distribution)

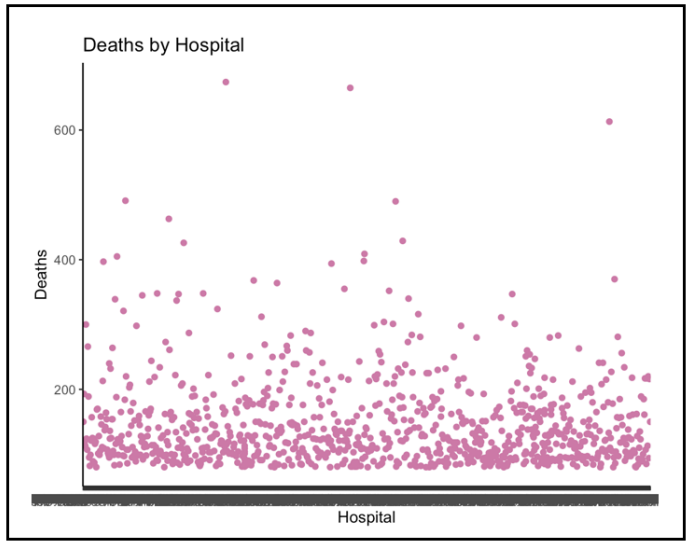


By Hospital

- ✓ Write a function

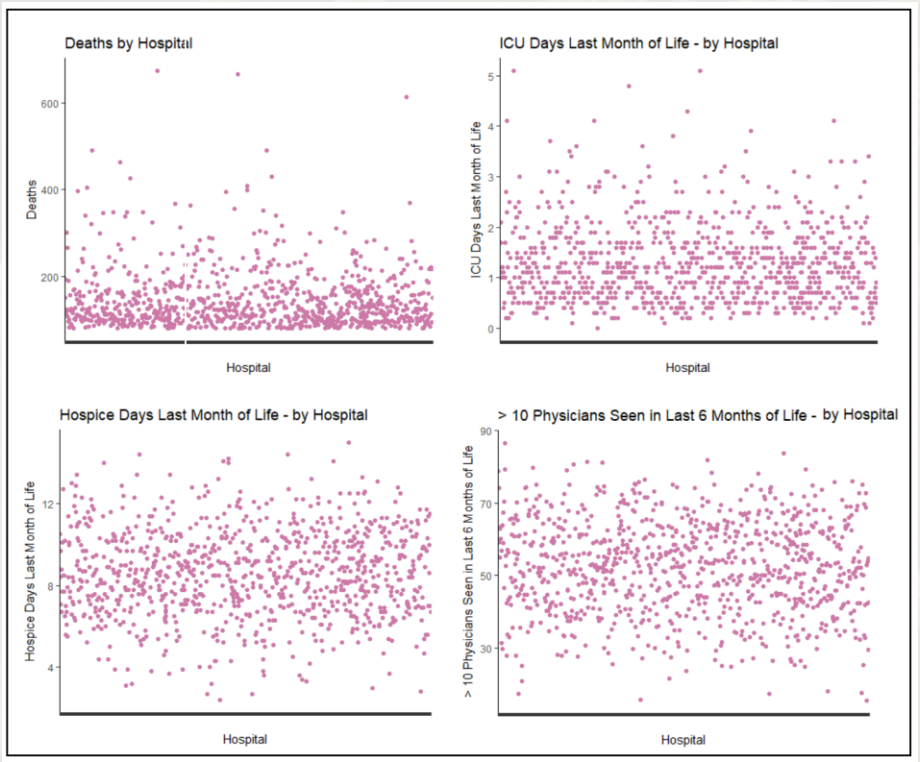
```
byhospital <- function(x){
  hospitalmap <- ggplot(DAPCancerEvents,aes(x=DAPCancerEvents$Hospital,y=col,colour=col))
  + geom_point(colour=cbbPalette[8]) +   ggtitle(title1)
  hospitalmap <- hospitalmap + labs(x=xlabel1,y=title)
  hospitalmap <- hospitalmap + theme_classic()
  hospitalmap
}
```

- ✓ Apply to each of our 12 data subsets



By Hospital

- ✓ Trends/Insights (using the same variables shown for histograms)
  - ✓ This data shows similar information as before, but the spread is shown along the y axis
  - ✓ The bottom two charts again look more normally distributed.

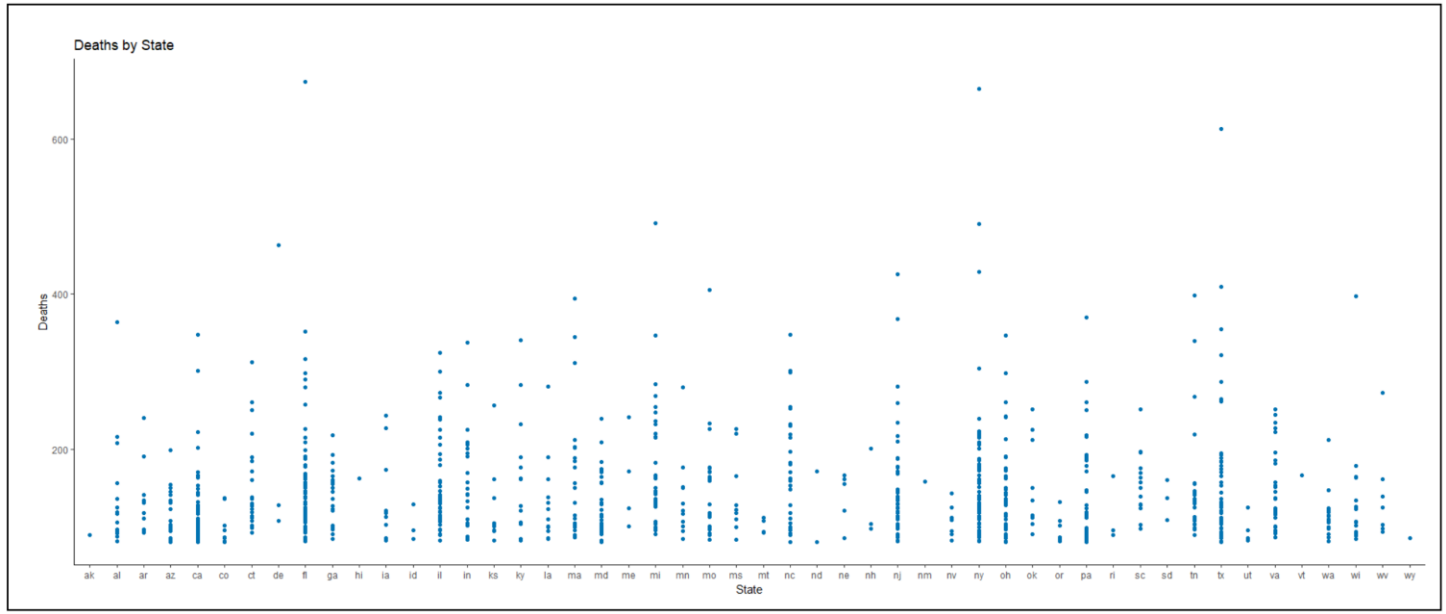


By State

- ✓ Write a function

```
bystate <- function(x){
  statemap <- ggplot(DAPCancerEvents,aes(x=DAPCancerEvents$state,y=col,
                                         colour=col)) + geom_point(colour=cbbPalette[6]) +
    ggtitle(title2)
  statemap <- statemap +labs(x=xlabel,y=title)
  statemap <- statemap + theme_classic()
  statemap
}
```

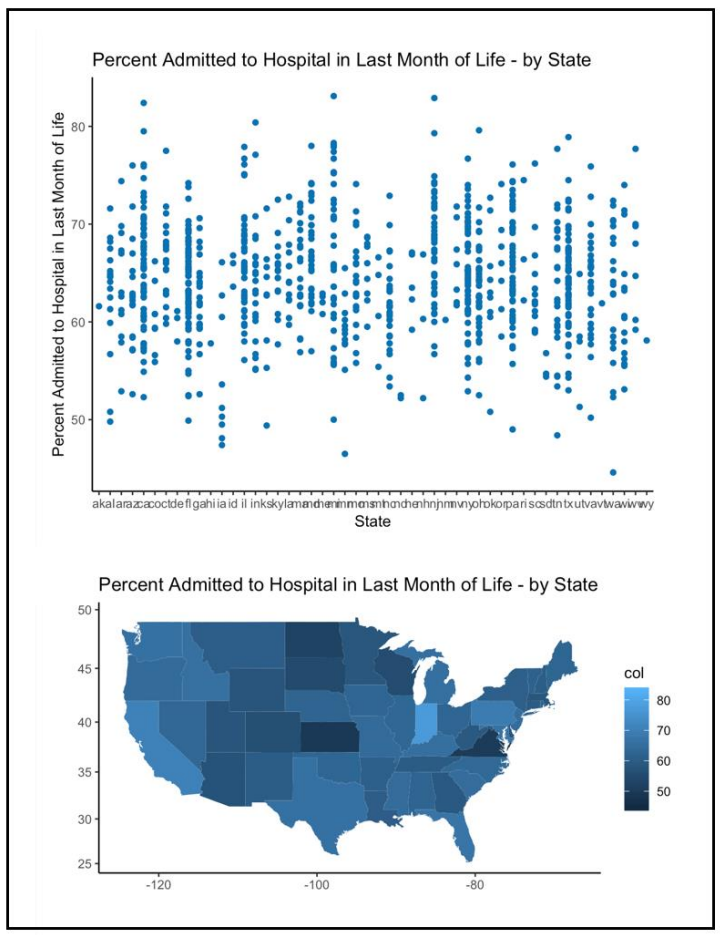
- ✓ Apply to each of our 12 data subsets



By State

- ✓ Trends/Insights:

  I think we learned that using geom_map to map the data out was a more effective way to visualize this data.

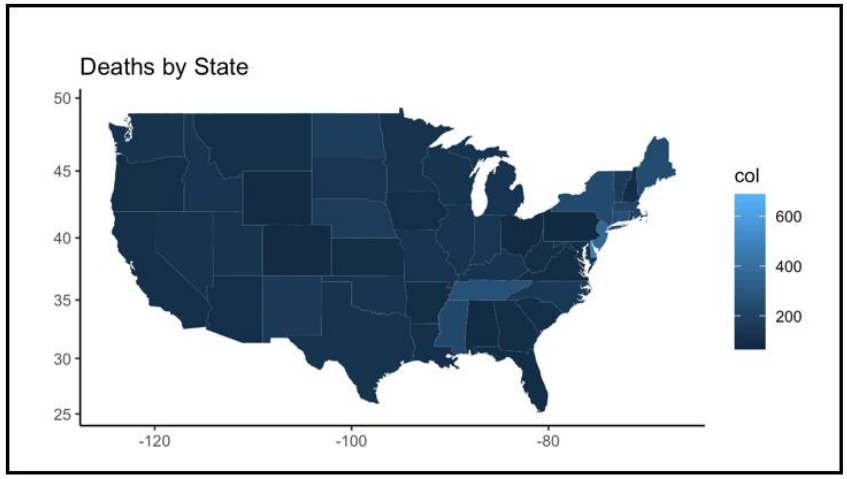


Visualize on US Map

- ✓ Write a function

```
#create function to generate maps
mapvariable <- function(x){
  us <- map_data("state")
  map <- ggplot(DAPCancerEvents,aes(map_id=statename))
  map <- map + geom_map(map=us,aes(fill=col))
  map <- map + labs(x=NULL, y=NULL)
  map <- map + expand_limits(x=
                             us$long, y=us$lat)
  map <- map + coord_map()
  map <- map + ggtitle(title2)
  map <- map + theme_classic()
  map
}
```

- ✓ Apply to each of our 12 data subsets



Visualize on US Map

- ✓ Trends/Insights:

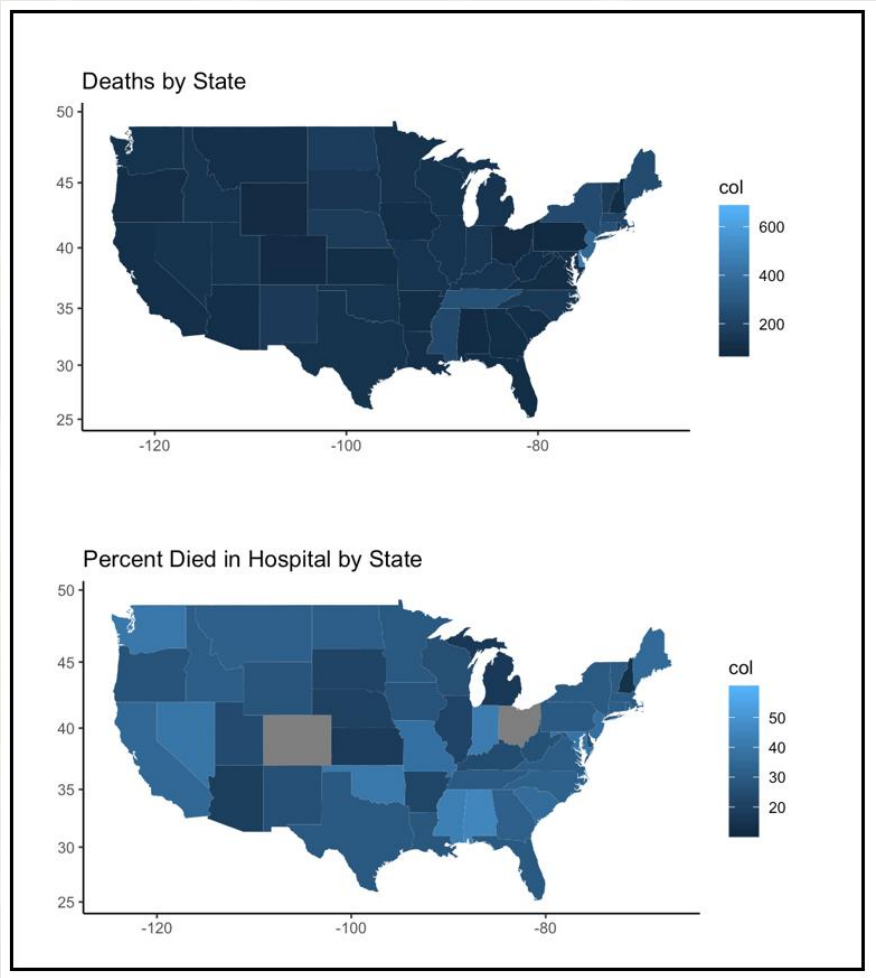
  Number of Deaths by State

  did not show much variation

  contrasted with

  Percent Died in Hospital by State

  (some data is missing)



Visualize on US Map

- ✓ Trends/Insights

  Hospice Days Last Month of Life &

  Percent Admitted to Hospital Last Month

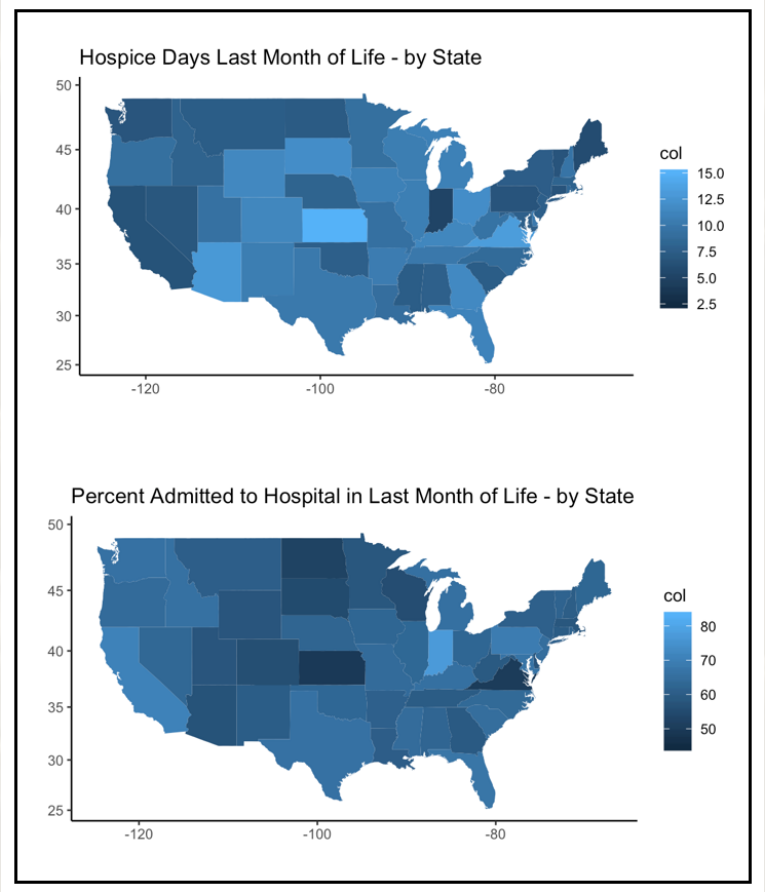
  Both showed variation,

  but it appears to be inverse.

  This makes sense, given the

  correlation of **-0.67**.

  (more on correlation next)



Correlation

- ✓ Run a correlation on 12 variables (columns 5-16)

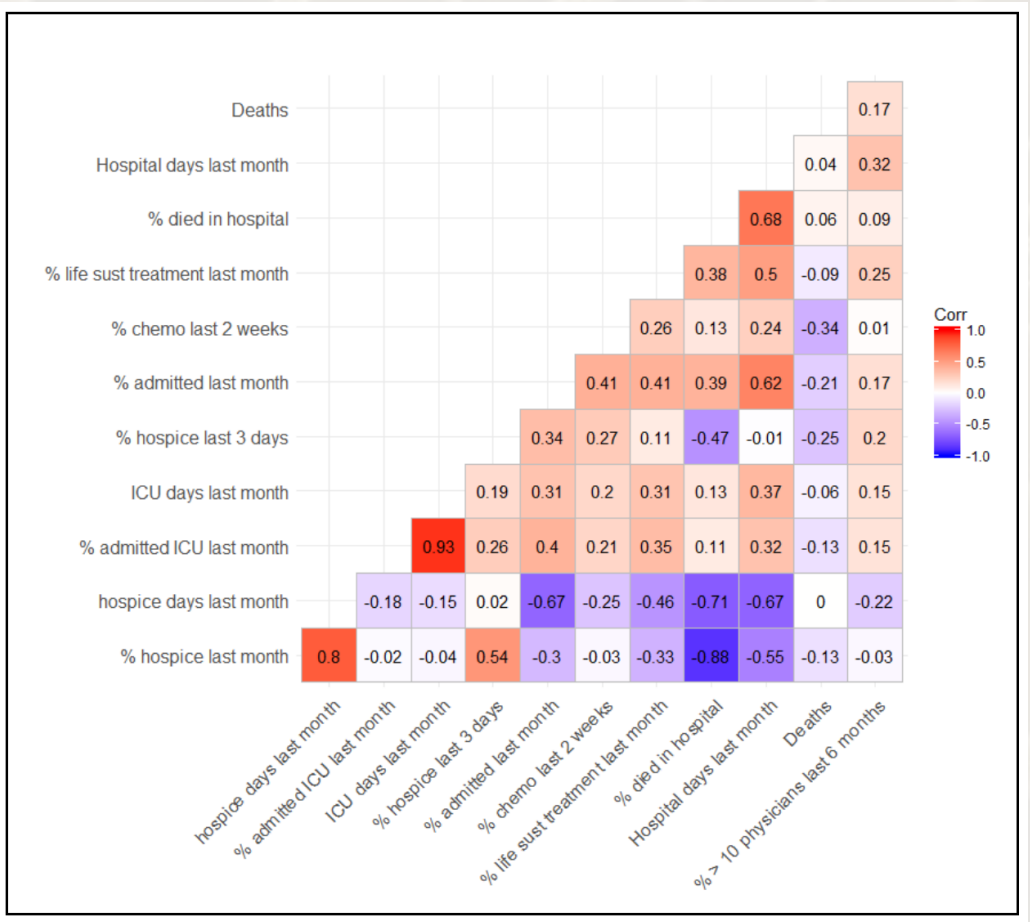
```
correlationdata <- DAPCancerEvents[,5:16]
DAP.cor<- round(cor(correlationdata,use="complete.obs"),4)
```

- ✓ Visualize using ggcorrplot package (visual on next slide)

```
install.packages("ggcorrplot")
library("ggcorrplot", lib.loc="~/R/win-library/3.4")
ggcorrplot(DAP.cor, hc.order=TRUE, type="lower", lab=TRUE)
```

- ✓ Strongest Positive Correlations:
  - ✓ % admitted ICU last month & ICU days last month = 0.93
  - ✓ % hospice last month & hospice days last month = 0.80
- ✓ Strongest Negative Correlations:
  - ✓ % hospice last month & % died in hospital = -0.88
  - ✓ hospice days last month & % died in hospital = -0.71

Correlation



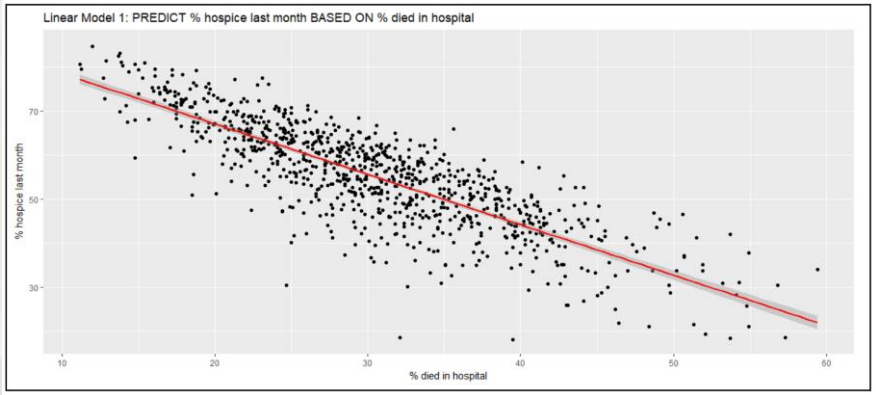
Linear Regression

- ✓ Wrote code for linear regression models

```
DAPCancerEvents.lm = lm(formula=DAPCancerEvents$`% died in hospital`~DAPCancerEvents$`hospice days last month`,
                        data=DAPCancerEvents)
DAPCancerEvents.model <- summary(DAPCancerEvents.lm)
paste ("Adjusted r squared=", DAPCancerEvents.model$adj.r.squared)
```

- ✓ Wrote code for visualizing models

```
g <- ggplot(DAPCancerEvents,
            aes(x= DAPCancerEvents$`% died in hospital`,y= DAPCancerEvents$`% hospice last month`))
g <- g + geom_point()
g <- g + stat_smooth(method="lm", col = "red")
g <- g + ggtitle("Linear Model 1: PREDICT % hospice last month BASED ON % died in hospital")
g <- g+labs(x="% died in hospital", y="% hospice last month")
g
```

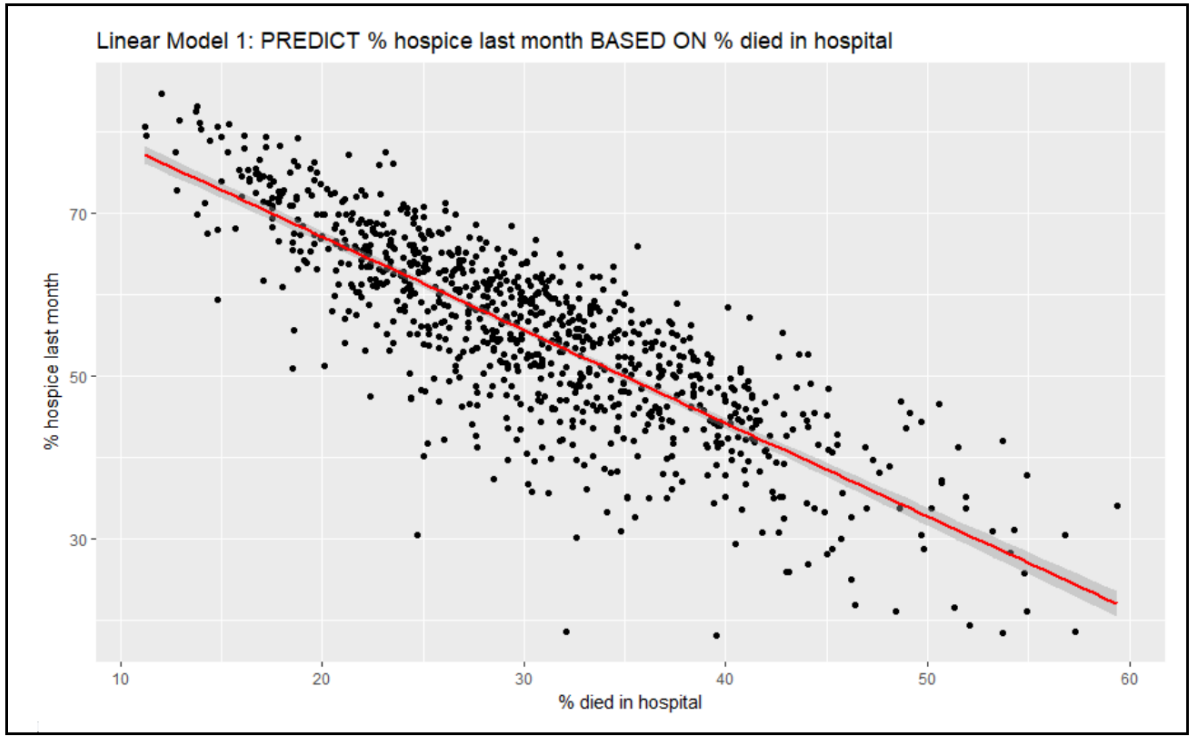


Linear Regression

- ✓ Run linear regression for all models with correlation stronger than 0.5 or -0.5
  - ✓ All P-values were significant
  - ✓ Adjusted R Squared ranged from 0.27 to 0.83
- ✓ Run linear regression on a model with multiple predictors
  - ✓ All P-values were significant
  - ✓ Adjusted R-Squared was 0.76
- ✓ Run linear regression on % died in hospital per state, then by city, then by both city & state
  - ✓ Most p-values were not significant
  - ✓ Including city & state gave the strongest Adjusted R-squared (0.51)

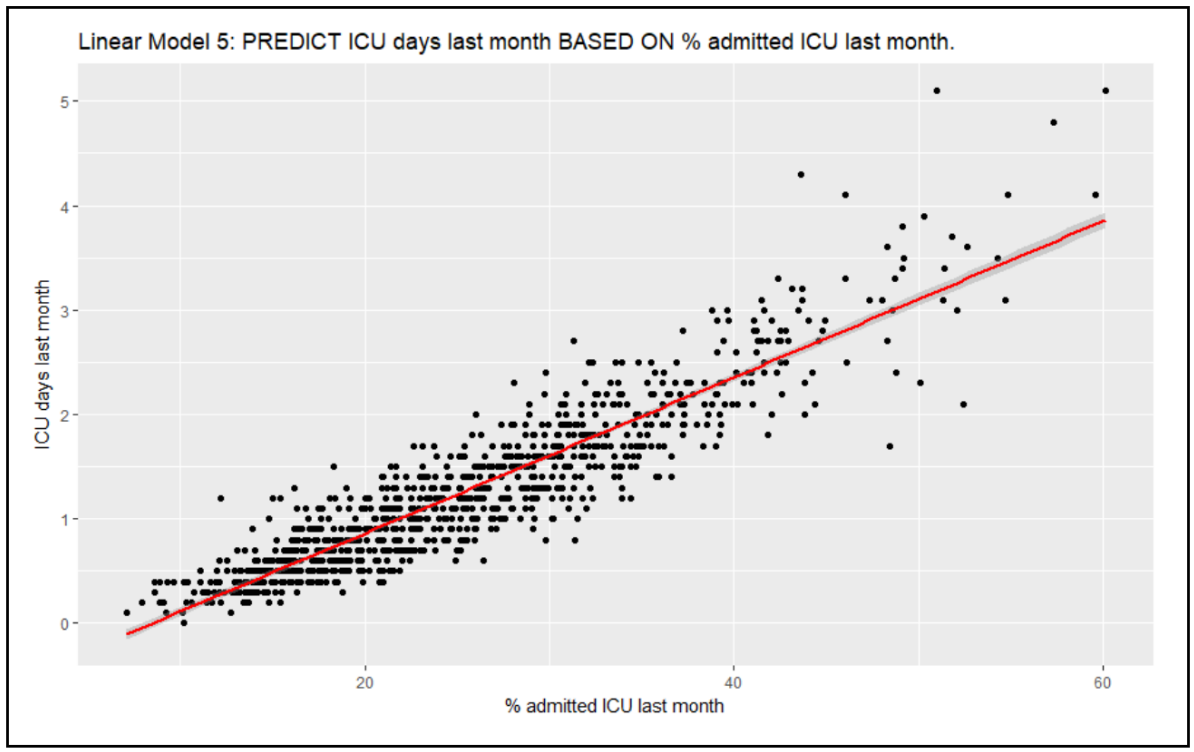
Linear Regression

- ✓ Model of strongest negative correlation:
  - ✓ Correlation was -0.88
  - ✓ Adjusted R squared = 0.67



Linear Regression

- ✓ Model of strongest negative correlation:
  - ✓ Correlation was 0.93
  - ✓ Adjusted R squared = 0.83



Conclusion

- ✓ **More is not always better**

EX: if we see our correlation chart, death is more negatively correlated with percentage of chemo last 2 weeks of life than percentage of hospice last month, what that means is even thou both of them are negatively correlated with death, its better to be in hospice care than in an aggressive treatment like chemo in the last weeks of life.

- ✓ **For Dying Cancer Patients, Geography is sometimes Destiny**

When we look at the end of life cancer care across geography and Hospitals, there are significant variations for different attributes

EX: Hospice days last month of life is more in Kansas, Virginia & Arizona on the other hand Hospital days last month of life is high in Oklahoma, Pennsylvania & Indiana. Also, If we take a look at percentage of life sustaining treatment last month of life, it is higher in eastern part of the country than western part of the country. Professionals may have to find out the why. So where you live and which hospital you go as a cancer patient decides what kind of care you will be getting.

- ✓ **Cancer is personal, so is the care patients need**

If we take a look at percentage of chemo last 2 weeks of life through geography and from hospital to hospital, its different because cancer is personal and the care patients need is different. Ex. Chemo might work well for one patient and not for another.

Also if we take a look at Hospice days and Hospital days last month of life, they are different across geography why? Because Cancer is personal and so is the care you need. Some patients might prefer hospice over Hospital and some might choose Hospital over hospice.

- ✓ **Aggressive Care for Dying Cancer Patients at end of life is hard and so its better to focus on keeping patients comfortable**

The data we are looking is poor prognosis cancer data, which means death is inevitable. If we again take a look at chemo in the last 2 weeks of life its negatively correlated with death, same is true for percentage of life sustaining treatment in last month of life. Since death is inevitable these aggressive treatments should be reduced and the focus might have to be on keeping patients comfortable

Reflections

- ✓ A couple of the columns in the data set were very sparse, and that was a bit disappointing. The maps for those columns had a lot of missing states.
- ✓ It was challenging to work with so many variables (12) that had such similar names (hospice sounds and looks a lot like hospital). It was easy to get things shifted/swapped, so we had to carefully double check our work.
- ✓ Initially, our descriptive statistics looked off, but we were able to solve it by utilizing the Numberize function.
- ✓ The Correlation didn't work well at first (there were a lot of N/As). This was resolved by replacing the code use="everything" with use="complete.obs"
- ✓ The correlation data was difficult to read, but we discovered the ggcorrplot package and it was an easy fix.