IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

**information technology & management**

# 529 Data Analytics

October 19th, 2016

Analysis on 2008–2010 DE-SynPUF

**Submission Document**

Submitted By
Naveed Ul Haq
A20357084

# Table of Content



ITM - 529

# Business Scenario
## Inpatient and Medicare Claims Data Study



**Inpatients and Benefit Claims**


- Medicare is a healthcare product and is a vital Government Policy
- The analysis on data of inpatients(Patients Admitted) treated at the healthcare facility and the subsequent Medicare claims is a major source of healthcare information
- Advancements in electronic information interchange of healthcare data and explosion of analytics in this field has made it feasible to study patterns and findings to augment the patient recovery rate and reduce risk areas

# Keywords/Glossary
## Inpatient and Medicare Claims Data Study



- Below listed are the Healthcare keywords/Terms used in this project
  - **Inpatients:** Patients **admitted to the healthcare facility (e.g. hospital)**
  - **CLM_PMT_AMT:** The **Claim Payment Amount** paid by the Medicare to the Healthcare Provider towards the service availed by the patient
  - **DESYNPUF_ID:** Surrogate Key which joins Inpatient File and Claims File
  - **DEATH_IND:** Binary Variable added to indicate if the patient death occurs in the hospital. Inpatients with death date not null add to mortality.
  - **LENGTH_STAY:** Duration of Stay of Inpatients in the healthcare facility. Difference between Claim Start Date and Claim End Date

# Business Objective
## Craft Data Story per



- The project is aimed at identifying health care related dependent and predictor attributes and modelling relationships between them
- The goal of this Project is to Build a Regression Model to predict Dependent **Claim Payment Amount** and **Death Indicator**
- **Predictors:**
  - **Age, Length of Stay**
  - **Ethnicity, Gender**
  - **Pre Ailments - Disease Indicator (Cancer, Heart Failure, Diabetes etc.)**
- Modelling decision tree to prove the hypothesis that a certain branch would have distinction in predicted attribute like **death indicator** based on different classifications of combination of dependent variables.

# Details of Data Sample
## 2008-2010 DE-SynPUF

- PUFs data is synthetic data maintained by Medicare and Medicaid Centers: https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DE_Syn_PUF.html
- **The entire population of data is sampled into 20 separate files, for ease of downloading the files, as the data is huge.**
- The website stores Claims, Inpatient and Outpatient Data along with the other files available for public use.
- Disclaimer: This analysis is only for the purpose of research as part of educational program. The sample and population of data is presented to agree to the terms and condition of data usage.

# Data Flow/Processing Steps
## Preparation Details of Data Solution



- The sample '**DE1_0_2008_to_2010_Inpatient_Claims_Sample_<1-20>.csv**' and '**DE1_0_2009_Beneficiary_Summary_File_Sample_<1-20>.csv**', each represent a 5% sample of the CMS PUFs synthetic data.
- The two sets of files are merged by the common field '**DESYNPUF_ID**' and contain 112 columns and 657586 rows upon merge
- The input file is imported in R and SAS Applications as data.frame and SAS7BDAT files respectively. Reporting, Visualization in Tableau & Excel
- The csv file prepared upon merge is called DeSynPUF_Claim_Inpatient.csv. It is **Merged, DeSynPUF, Med_Desynpuf** in different iterartions
- **Columns Added: Age, Death-Ind, Length_Stay**

# Project Overview
## Tool, Technology and Templates



- **Platform/Language:** R Studio 3.3.2, SAS Community Edition, Tableau
- **SAS Procs:** Means, Freq, SQL
- **CRAN Packages/Libraries:** rpart, rattle, tree, logicForest, party, ggplot2, plyr,
- **Aim:** Build a prediction model of survival for inpatients.
- **Data Governance:** Masking Compliance as per HIPAA Privacy Rule
- **Model:** Regressions, Predictions and Decision trees
- **Statistical Method:** Machine Learning - random forest/recursive partitioning
- **Learning:** Unsupervised
- **Approach:** Iterative
- **Predictors (X):** Gender, age, reason of admission, length of stay and pre-conditions, ailments, health flags
- **Modeled Dependent Variable (Y):** A binary dependent Policy Holder Retention Indicator

# Project Plan/Data Activities
## Stages of the Project

- **Design:**
  - Data considered only - LA County Crime, Consumer Complaints
  - Basis of Selection - Rich, Clean Data with many attributes
- **Model Based / Data Driven:**
  - Choice of Regression Model – Logistic or Multi-variate
  - Based on the types of variables - Binary, Continuous or Categorical
- **Development:**
  - The R Script and SAS Code is written, tested and output noted
- **Result:** Presentation, Visualization and Documentation

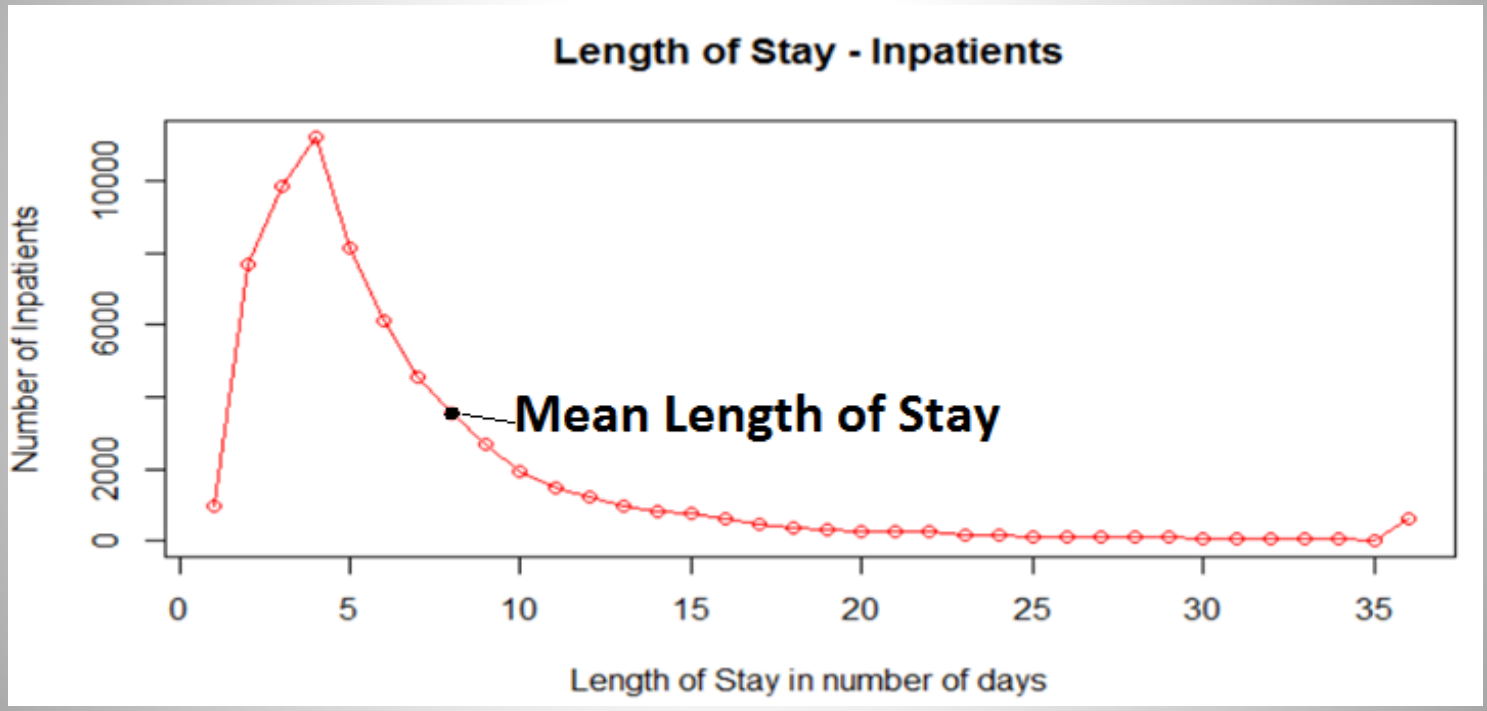
# Exploratory Data Analysis Results
## Medical Conditions



- **Major Diseases**
  The most common major disease is Stroke and Cancer followed by Osteoporosis and Arthritis

# Exploratory Data Analysis Results
## Patient Count vs Length of Stay in Hospital

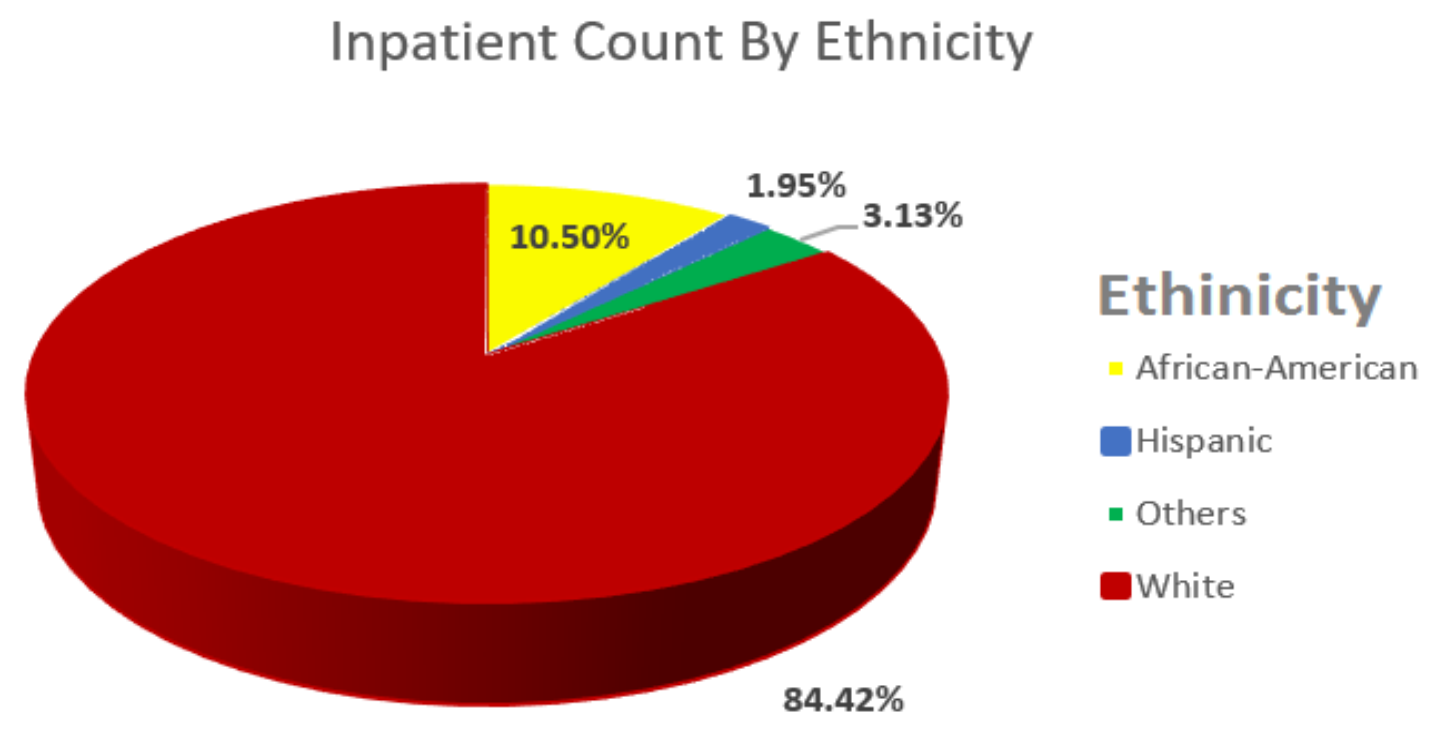
- **Length of Stay of Patients**
  Most of the patients stay less than 10 days in hospital. The **mean** Length of Stay is **7 Days**.



# Exploratory Data Analysis Results
## Patient Count By Ethnicity

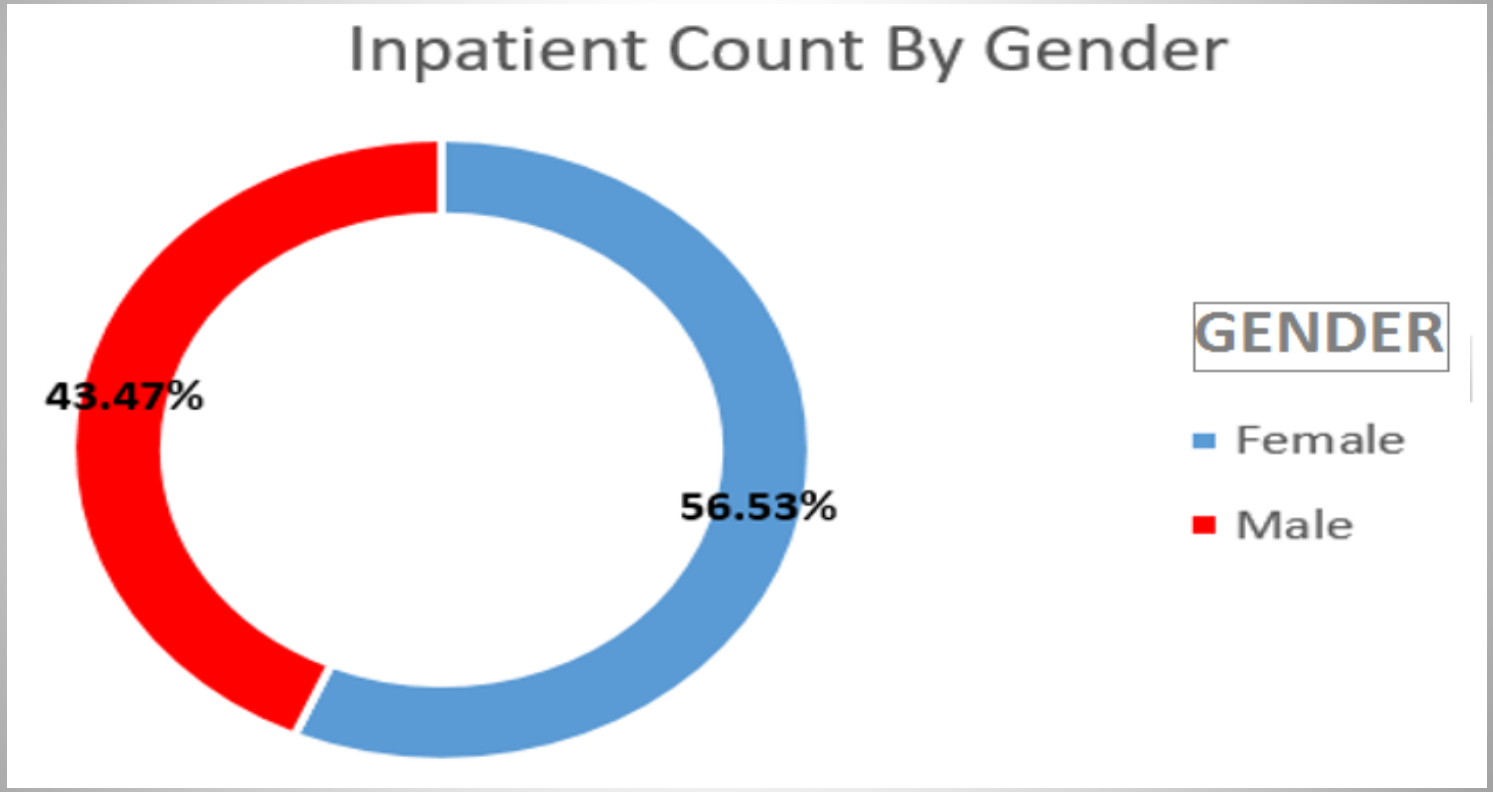


- **Inpatient Count Summary by Ethnicity**
  Hispanic constitute the least, whereas African American population is substantial



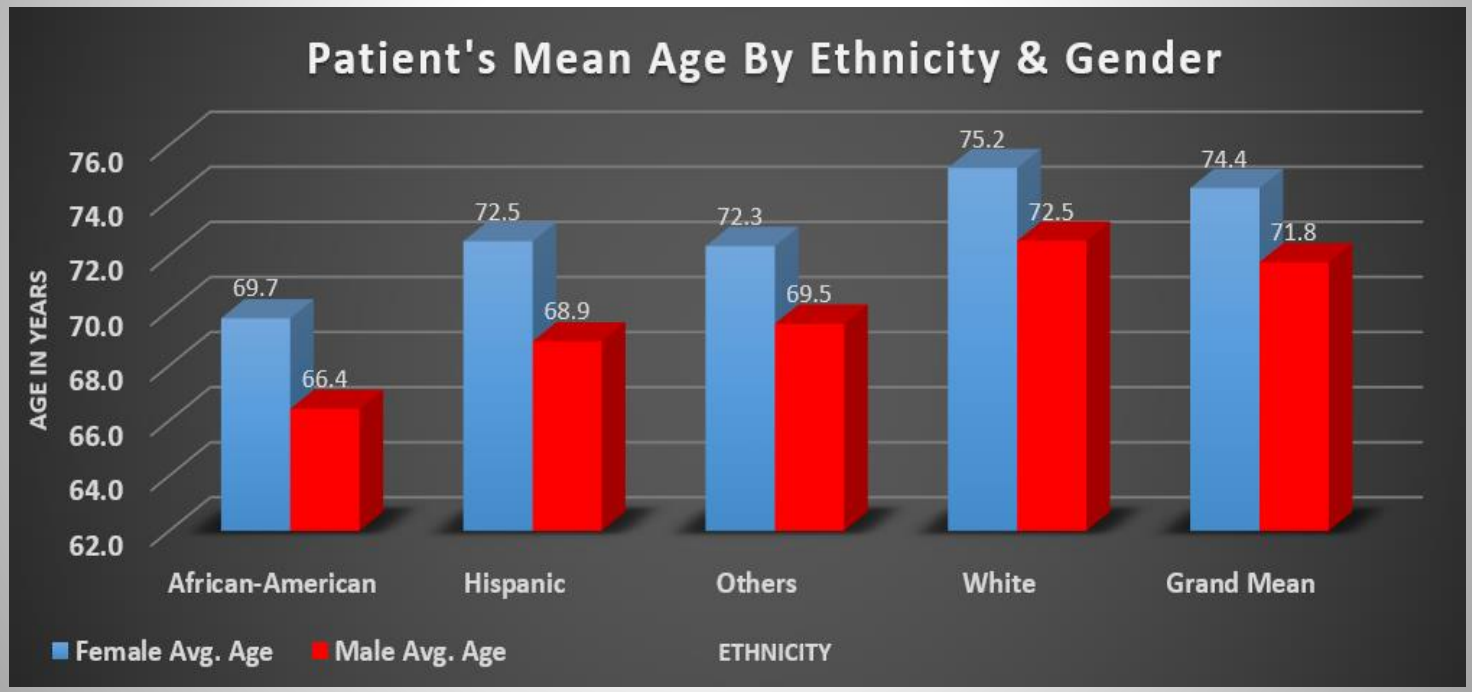
# Exploratory Data Analysis Results
## Inpatient Count Summary

- **Inpatient Count Summary by Gender**
  More Female are hospitalized than Male. The difference is greater than 13%



# Exploratory Data Analysis Results
## Inpatient's Mean Age

- **Inpatients Mean Age By Ethnicity and Gender**
  - The mean age of Inpatients for Male is lesser than the female
  - Mean age of African-American Inpatient is the least.
  - White Female are hospitalized at older age.



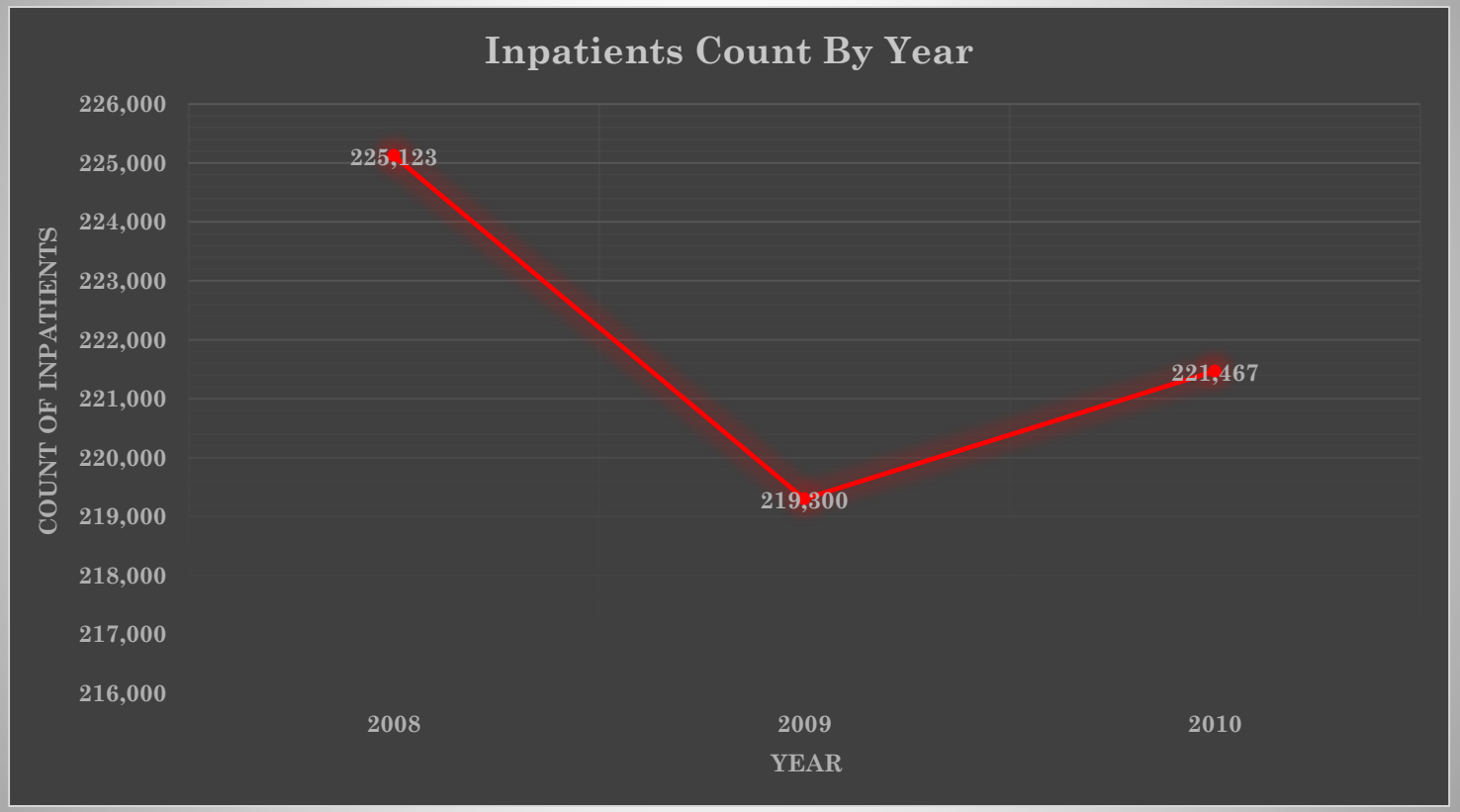

ITM - 529

# Exploratory Data Analysis Results
## Count Of Inpatients By Year over Year



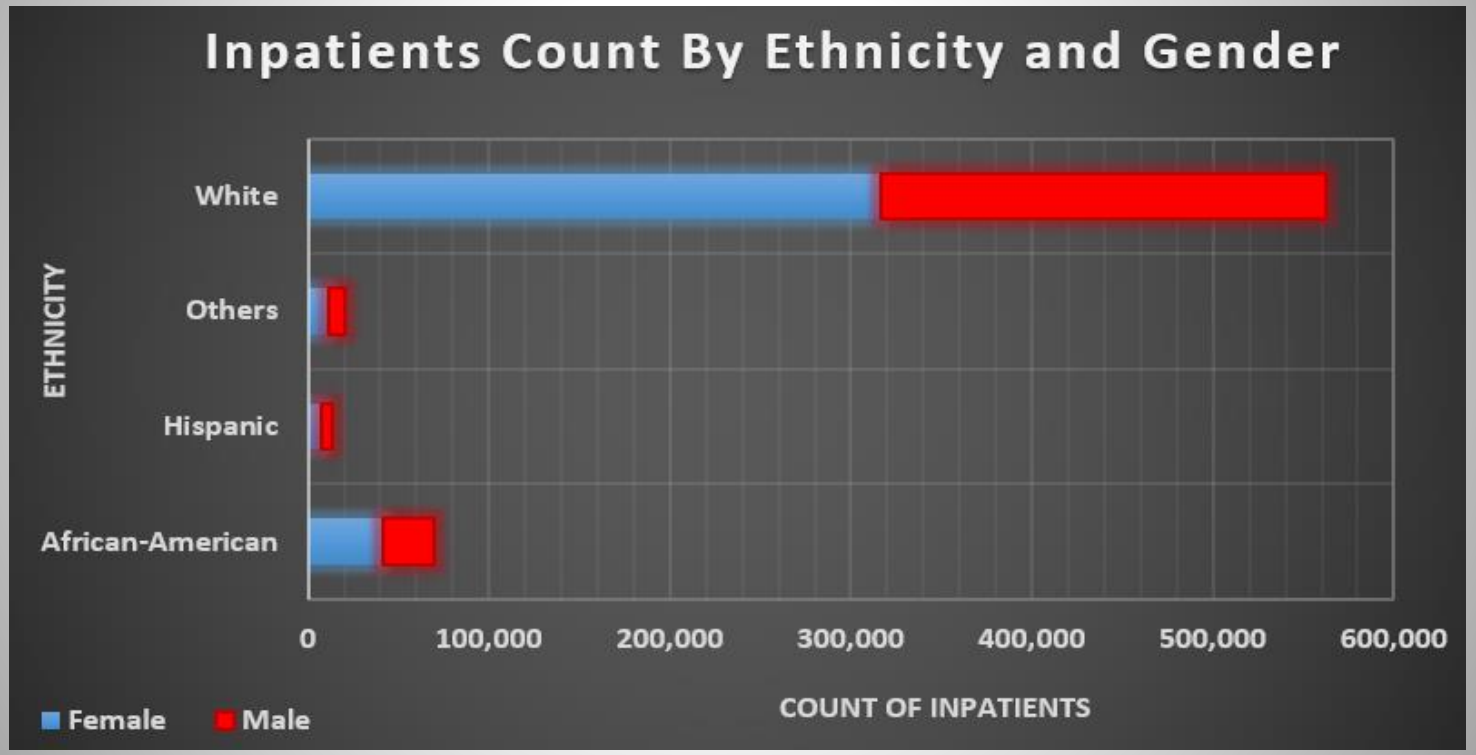
- **Inpatients Count**

  The Number of Inpatients witnessed a steep fall in 2009.



# Exploratory Data Analysis Results
## Inpatient's Count Summary

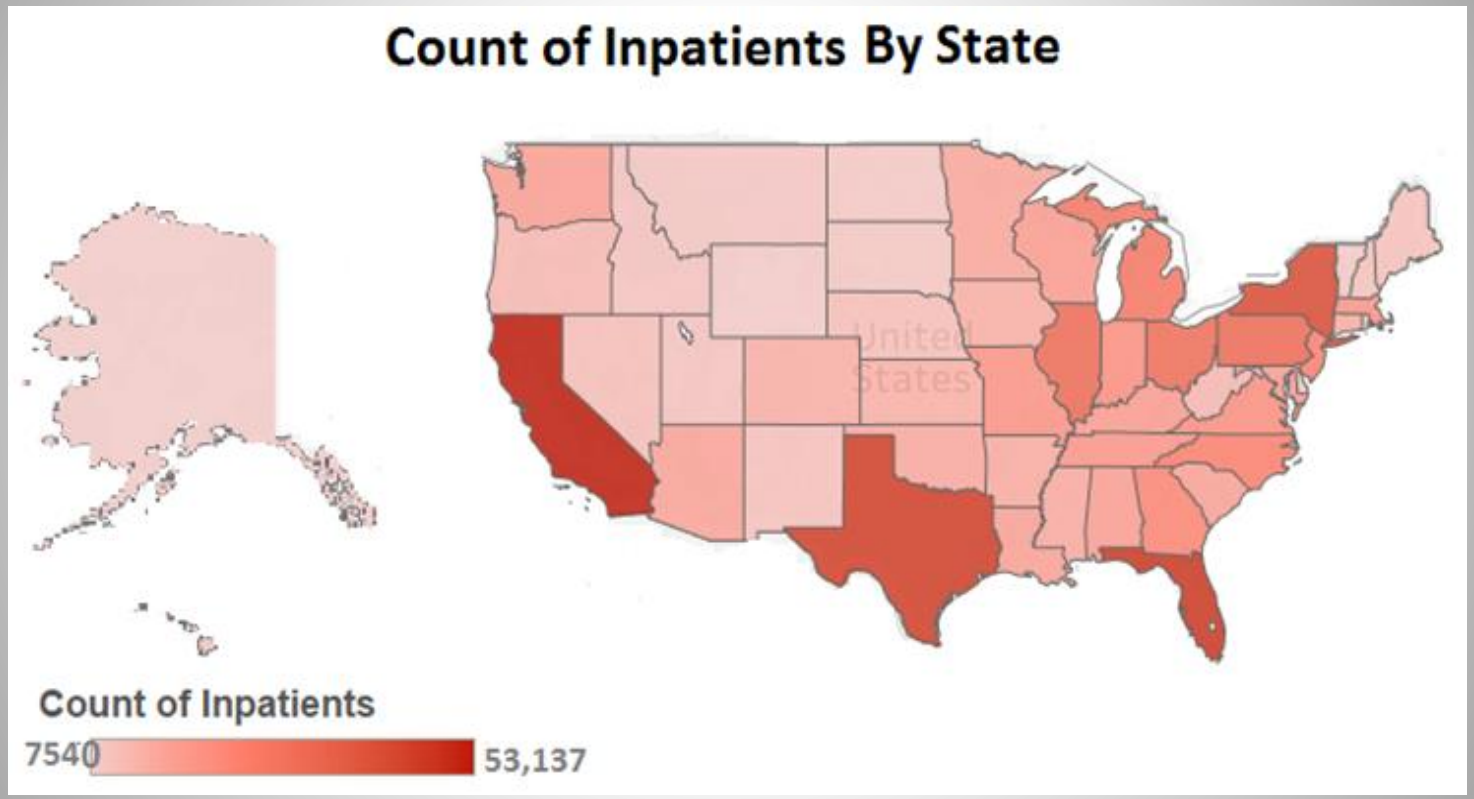
- **Inpatients Count By Ethnicity and Gender**
  Clearly more female are admitted as compared to the male population.
  African American are hospitalized more in proportion to their population.



# Exploratory Data Analysis Results
## Inpatient's Count By State



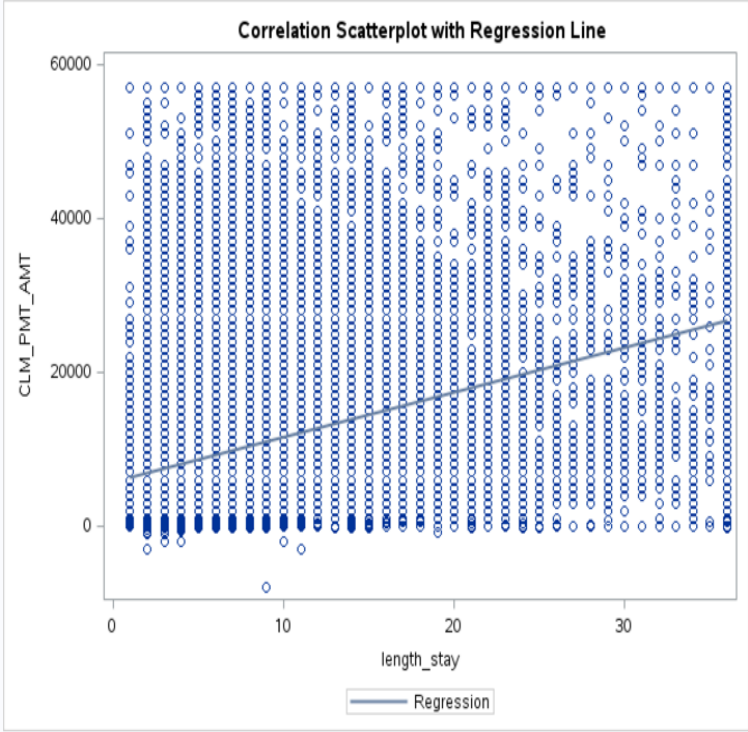
- **Inpatients Count summarized by State**
  California, Texas & Florida have highest number of Inpatients as per population



# Exploratory Data Analysis Results
## Correlational Scatterplot and Diagnostics



- **Claim Payment Amount, Length of Stay and Age**
  The Claim Payment Amount has positive correlation of 0.36 with length_Stay (Claim Start date Through Claim End Date)




A20357084 - Correlations with Claim Payment Amount

The CORR Procedure

| | |
|---|---|
| 1 With Variables: | CLM_PMT_AMT |
| 2 Variables: | AGE length_stay |

Simple Statistics

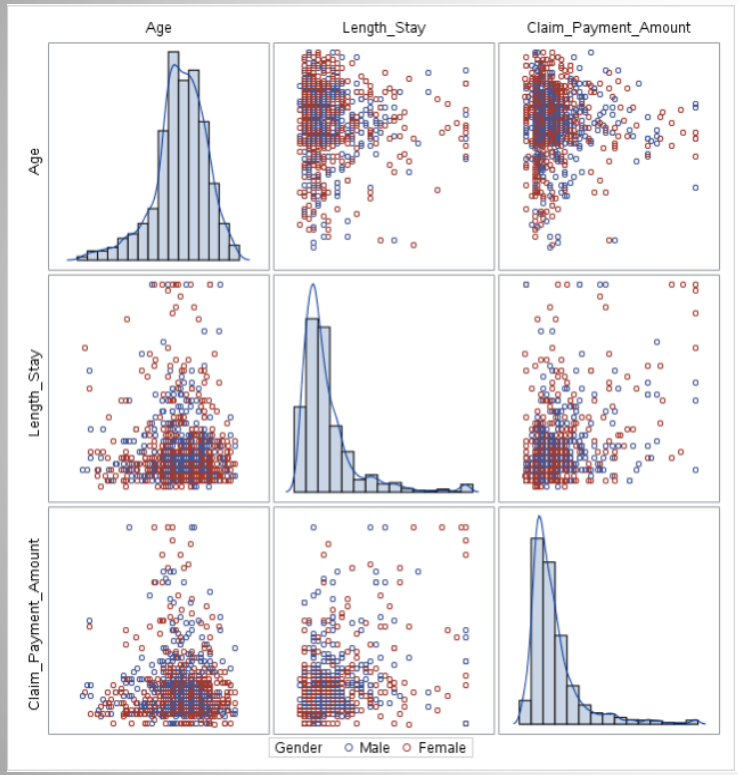
| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| CLM_PMT_AMT | 66586 | 9573 | 9312 | 637429980 | -8000 | 57000 |
| AGE | 66586 | 73.27572 | 13.26861 | 4879137 | 24.00000 | 101.00000 |
| length_stay | 66586 | 6.71426 | 5.73658 | 447076 | 1.00000 | 36.00000 |

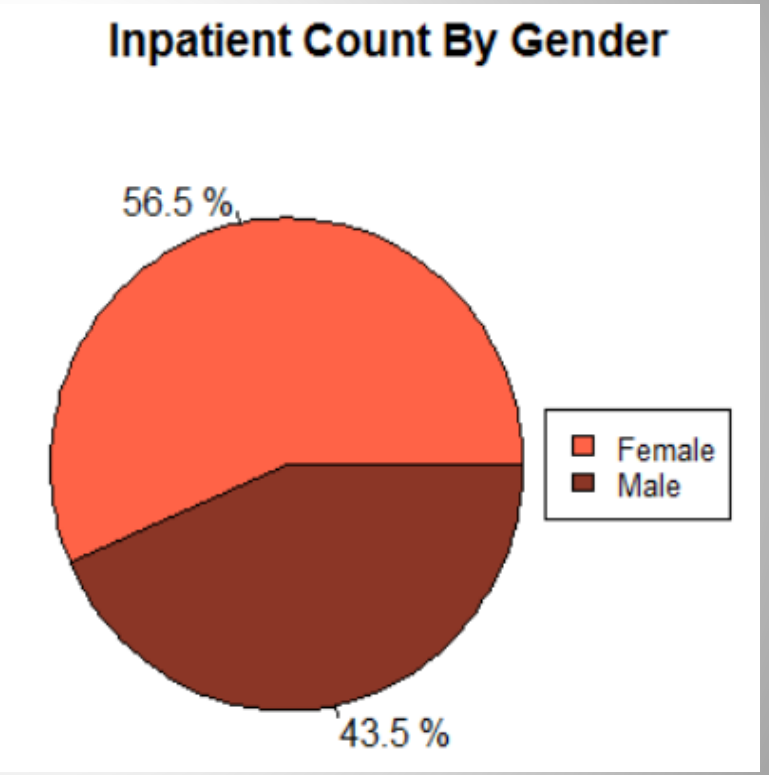
Pearson Correlation Coefficients, N = 66586
Prob > |r| under H0: Rho=0

| | length_stay | AGE |
|---|---|---|
| CLM_PMT_AMT | 0.36056<br><.0001 | 0.00535<br>0.1673 |

# Exploratory Data Analysis Results
## Scatterplot Matrix

- **Scatterplot/Correlation of variables**
  Continuous and Categorical Demographic Attribute plot. Pie Chart for insight

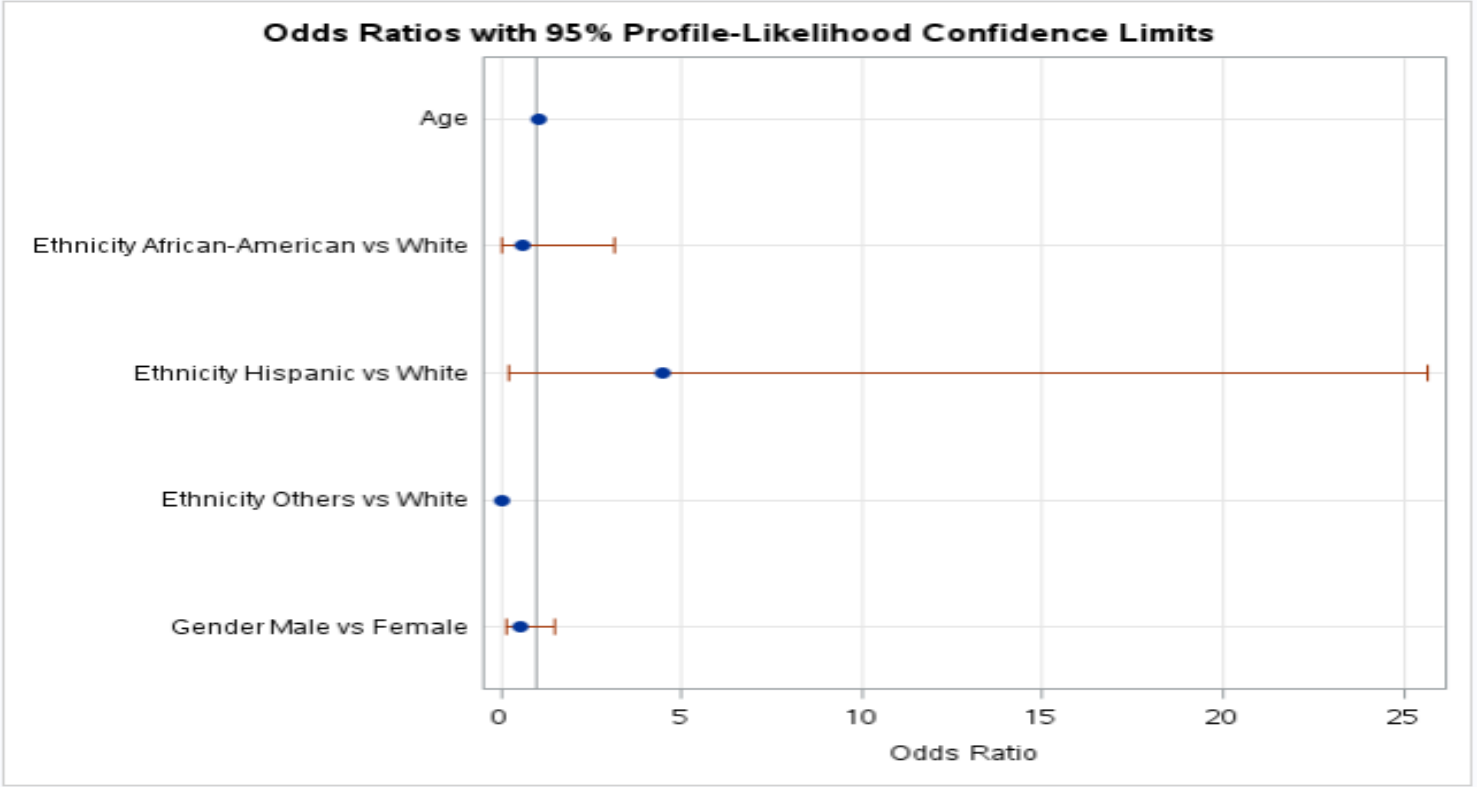



ITM - 529

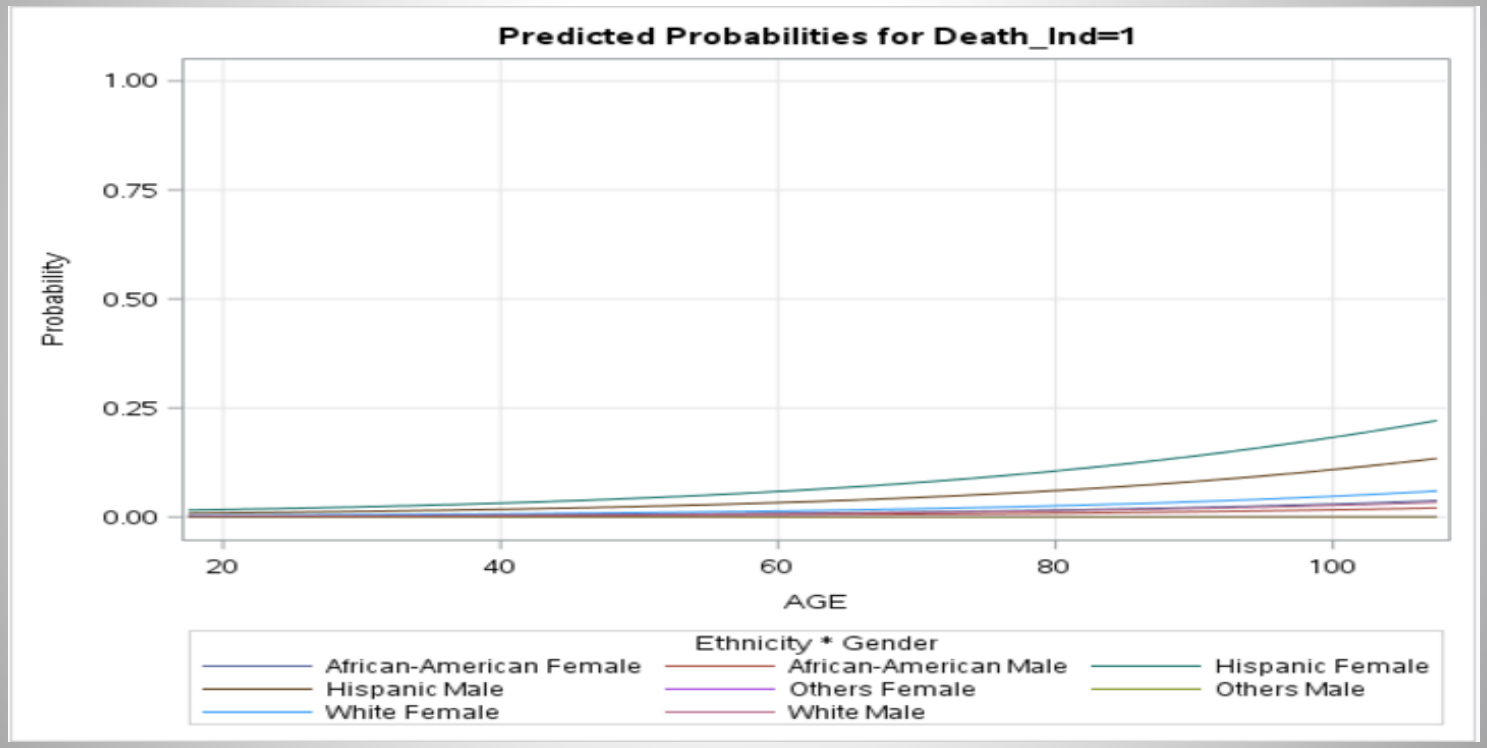
# Odds Ratio
## Ethnicity and Gender

- **Predicted Probability of Inpatient Deaths of Emergency**
  Evidently Higher Odds Ratio of **Hispanic** vs **White** ethnicity & high variance



# Regression Modeling Results
## Logistic Regression Plot

- **Predicted Probability of Inpatient Deaths of Emergency Inpatients**
  The Death Indicator is plotted by Age and categorical variables Ethnicity and Gender. Male and Female population have distinct behavior.



# Regression Modeling Results
## Claim Payment Amount - Best Fit Model



- **Linear Regression Model of Claim Payment Amount on continuous variables - Age and Length of Stay.**

| | | | |
|---|---|---|---|
| **Root MSE** | 8231.53535 | **R-Square** | 0.1654 |
| **Dependent Mean** | 9655.00000 | **Adj R-Sq** | 0.1637 |
| **Coeff Var** | 85.25671 | | |

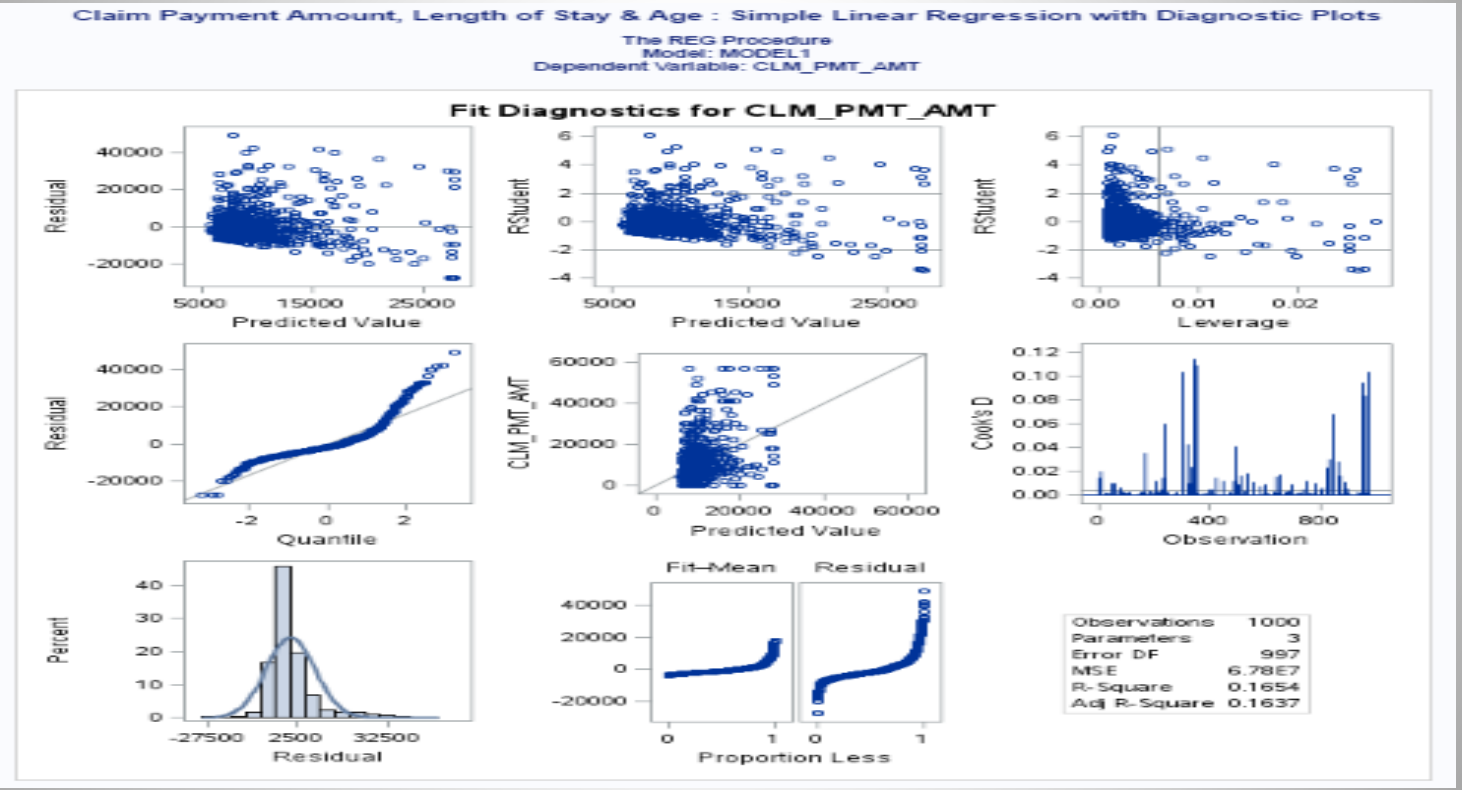
**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Standardized Estimate | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|---|
| **Intercept** | 1 | 6232.47445 | 1536.75258 | 4.06 | <.0001 | 0 | 3216.83381 | 9248.11510 |
| **length_stay** | 1 | 615.16508 | 43.86198 | 14.03 | <.0001 | 0.40594 | 529.09268 | 701.23748 |
| **AGE** | 1 | -11.12003 | 20.14489 | -0.55 | 0.5811 | -0.01598 | -50.65129 | 28.41122 |

# Regression Modeling Results
## Fit Diagnostics for Claim Payment Amount



- **Linear Regression with Diagnostic plots for Best Fit of Model**
  The **regressors** - Length of Stay and Age of patients are with residuals.



# Multivariate Regression
## Claim Payment Amount & Death Indicator



- **Dependents:** Age, Length_Stay, Ethnicity, gender, Cancer, Heart Failure Alzheimer, Kidney and Pulmonary Disease Indicator

The GLM Procedure
Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for AGE
E = Error SSCP Matrix

| Characteristic Root | Percent | Characteristic Vector V'EV=1 | |
|---|---|---|---|
| | | Death_Ind | CLM_PMT_AMT |
| 0.00005222 | 100.00 | 0.01704313 | 0.00000040 |
| 0.00000000 | 0.00 | 0.03619774 | -0.00000019 |

The GLM Procedure
Dependent Variable: CLM_PMT_AMT

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 765232298304 | 69566572573 | 924.65 | <.0001 |
| Error | 66574 | 5.0087514E12 | 75235848.405 | | |
| Corrected Total | 66585 | 5.7739837E12 | | | |

| R-Square | Coeff Var | Root MSE | CLM_PMT_AMT Mean |
|---|---|---|---|
| 0.132531 | 90.60723 | 8673.860 | 9573.033 |

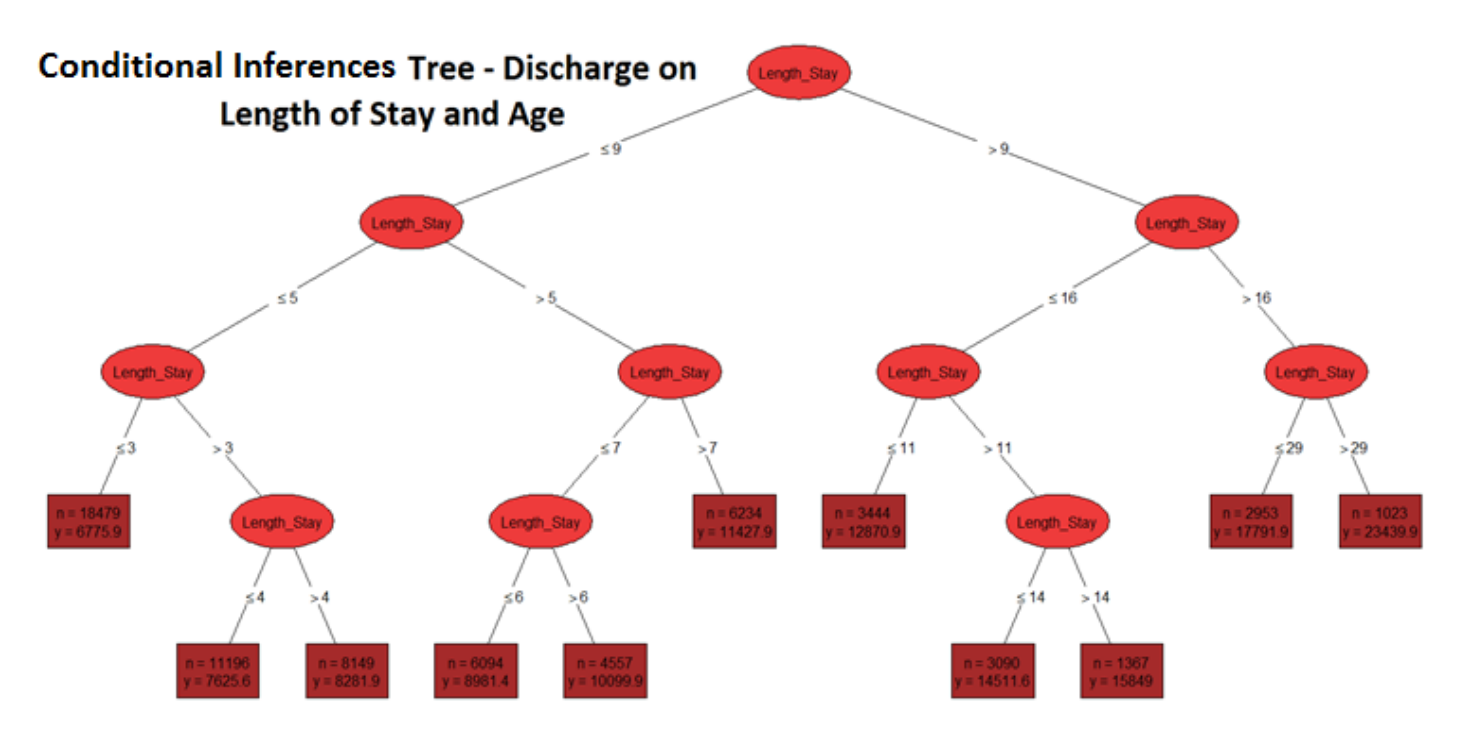
The GLM Procedure
Dependent Variable: Death_Ind

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 11 | 0.3083725 | 0.0280339 | 2.99 | 0.0006 |
| Error | 66574 | 624.7119771 | 0.0093837 | | |
| Corrected Total | 66585 | 625.0203496 | | | |

| R-Square | Coeff Var | Root MSE | Death_Ind Mean |
|---|---|---|---|
| 0.000493 | 1022.212 | 0.096870 | 0.009476 |

# Regression Modeling Results
## Conditional Inferences Tree of Discharges

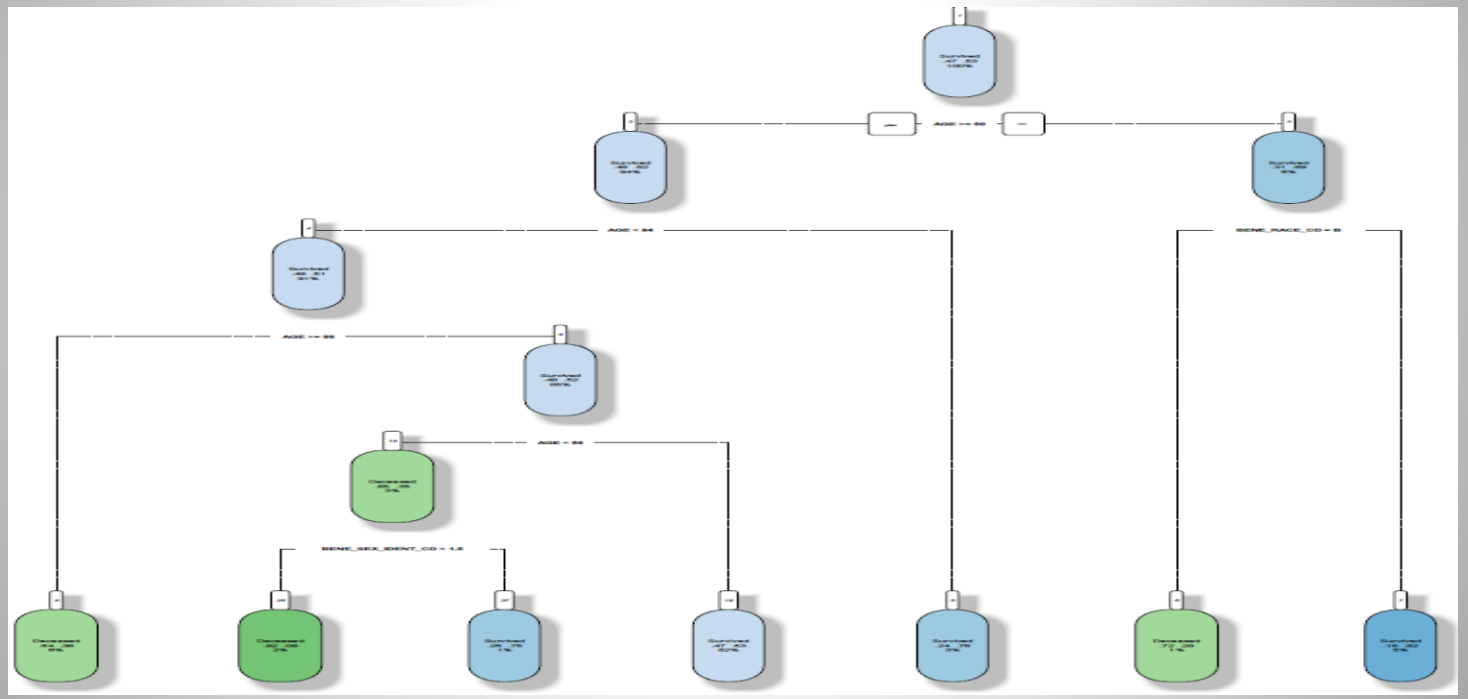
- **Critical Inpatients - Discharge Rate Partition Tree**
  The tree nodes propagate without pruning. Distinct classification on Length of Stay than on age. **N contribute to the Mortality and Y to Discharge**



# Recursive Partition Tree For Prediction
## Discharge of Critical Inpatients



- **Survival Prediction of Critical Patients:** Discharge Rate of Emergency patients on dependent variables – Age, Gender, Race, Length of Stay. African-American Male of age greater than **88 have only 2% survival**. **Random Forest(Machine Learning) & Rattle-Rpart Package.**



**Note: Visual is distorted as Presentation View Limited. Partition Tree Document attached as Submission Deliverable.**

# Summary and Conclusion

- From **2008-2010** on an average YoY 221,963 or 0.21 Million Inpatients
  - Female population is more in numbers and exceed male by 13.0%
  - The mean age of Female Inpatients is lesser than Male
- The mean age of African American Male availing Medicare Benefits is the least and claim amount/person do not vary much by state. White Female have **best health conditions** as their mean age is highest. But count of inpatients is also more in female.
- The mortality rate of critical inpatients is highest in Black African American Male of 80 years or more of age.
- **The Death of Inpatients is dependent sparsely on age and length of stay in hospital.**
- **Claim Payment Amount is not dependent only on age, length of stay & pre-ailment only. It varies with facility, services, reason of admit, insurance policy product and likewise.**

# End of Slide