DATA SCIENCE HACKATHON 2018

PREDICTING IDEAL LENGTH OF STAY FOR 30 DAY READMITTED VISITS

GROUP 2 – SAINT PETER'S UNIVERSITY
NANDINI KHATOD
PREETHI RAVI
PRANAY KATTA
VINOTHINI PUSHPARAJA
EBERLYNE LEGERME

OBJECTIVES

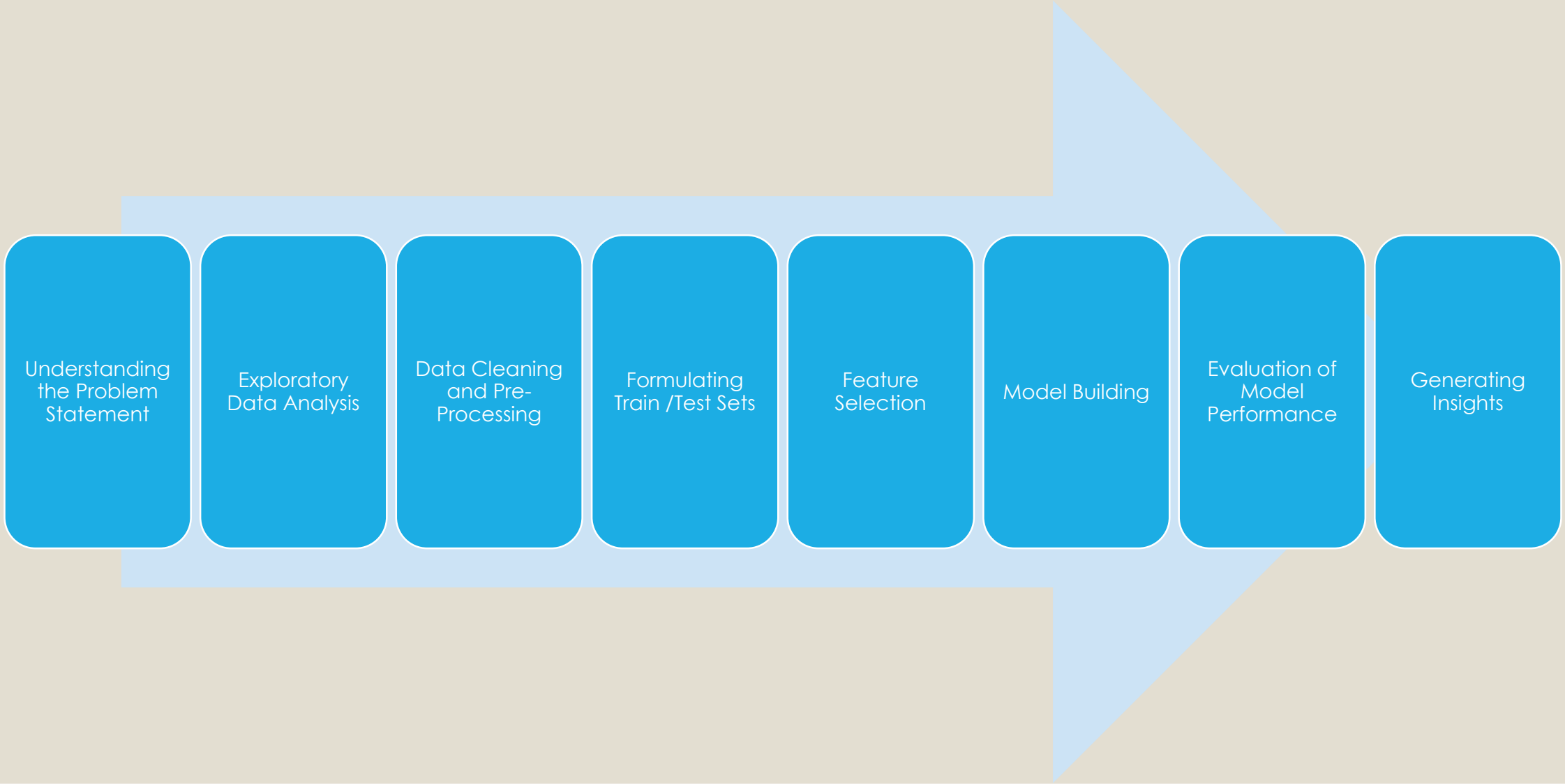
- Predicting the ideal length of stay for liver patients who got readmitted within first 30 days of their index admission
- Calculating cost saved per visitlink

Findings

Average increase in LOS for index VisitLinks = 2.95 ~ 3 days
Average Cost Saved per VisitLink = $56,523.94

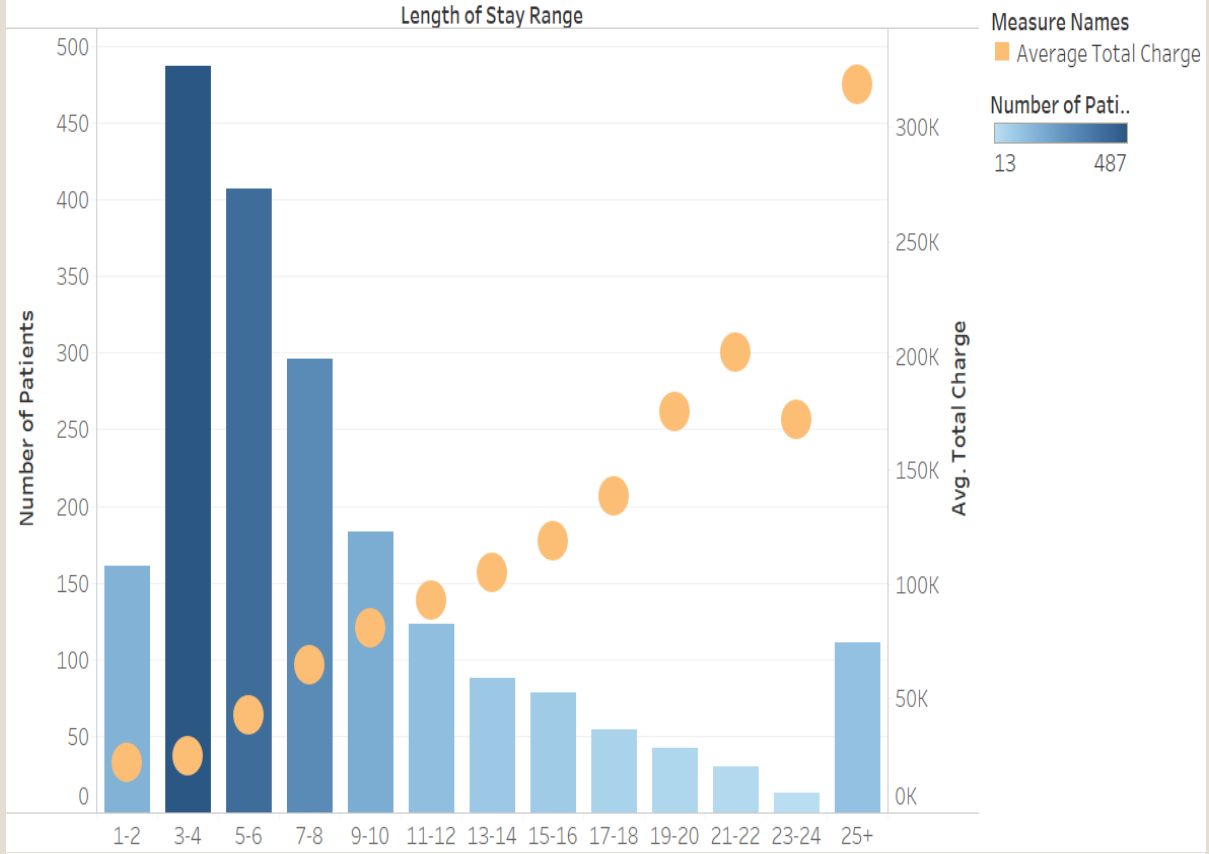
- By increasing average LOS by 2.95 or 3 days in first index visit $56,523.94 costs will be saved on an average per VisitLink.
- By staying 3 more days on the index visit, 3 extra days or 6 days in total of stay during readmission can be eliminated on an average
- High correlations are found between ADC (Average daily Census), ADMTOT (Total Facility Admissions), BDTOT (Total Beds Staffed)

Process flow

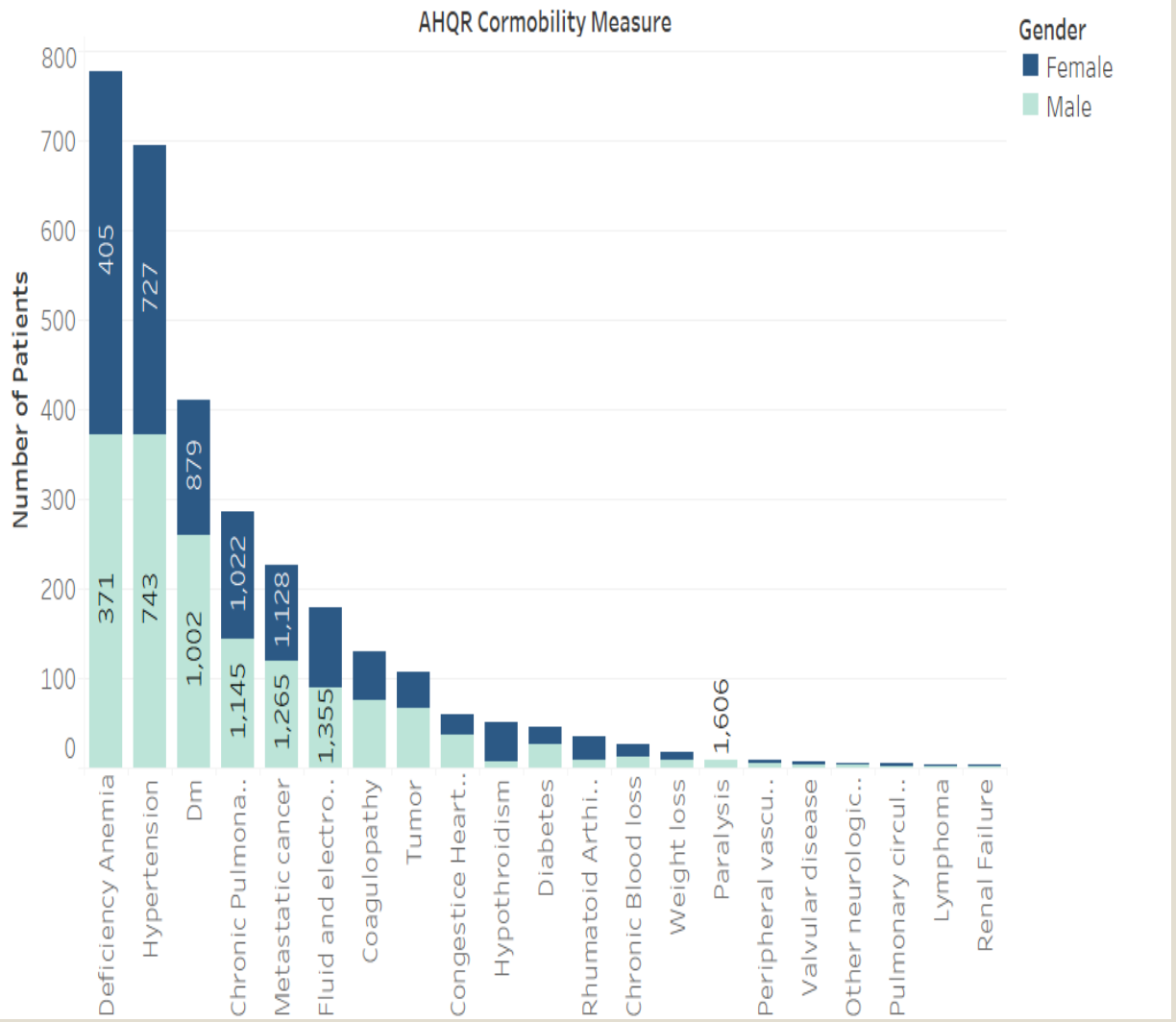


Exploratory data analysis

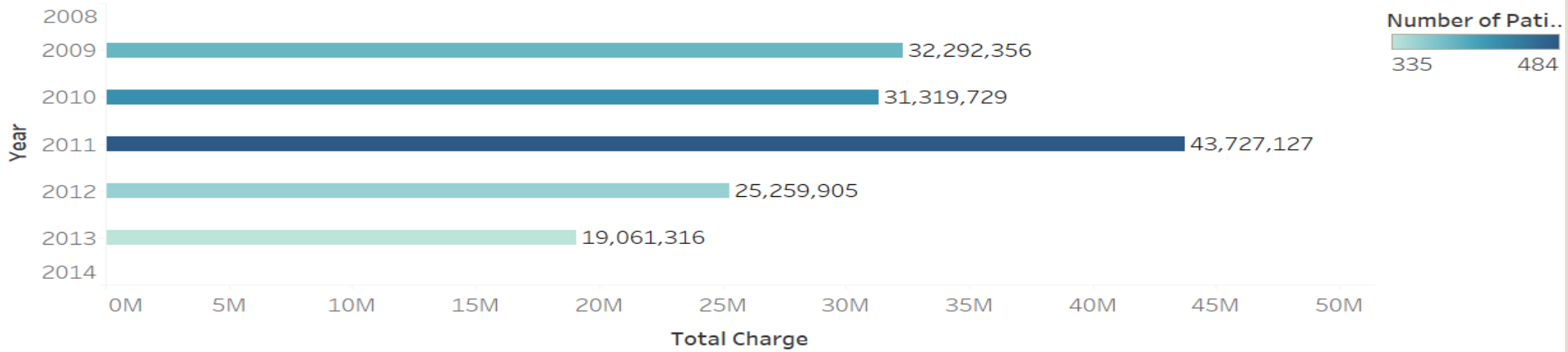
Average charges for the length of stay



The causes of readmission by gender



Yearly expense for readmission



Readmission type

Data cleaning and pre-processing

```
In [5]: print(df.isnull().sum())
ATYPE            0
AWEEKEND         0
BDTOT            0
CHC              0
CM_AIDS       1773
CM_ALCOHOL    1403
CM_ANEMDEF       0
CM_ARTH          0
CM_BLDLOSS       0
CM_CHF           0
CM_CHRNLUNG      0
CM_COAG          0
CM_DEPRESS    1403
CM_DM            0
CM_DMCX          0
CM_DRUG       1403
CM_HTN_C         0
CM_HYPOTHY       0
CM_LIVER      1403
CM_LYMPH         0
```

	ADC	ADMTOT	AGE	BDTOT	FTMDTF	FTRES	FTRNTF	NCHRONIC	NDX	NPR	SUF
count	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.000000	31283.0
mean	504.913371	32084.854426	58.755171	645.792827	174.708979	318.544577	1307.326695	4.402647	9.476009	3.386152	10482.9
std	413.937023	25137.223522	14.740317	508.495655	343.847205	407.436866	1117.885653	2.681588	5.252218	3.077413	8057.3
min	3.000000	192.000000	18.000000	16.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.0
25%	224.000000	16327.000000	50.000000	315.000000	2.000000	4.000000	487.000000	2.000000	6.000000	1.000000	4329.0
50%	386.000000	24515.000000	60.000000	470.000000	23.000000	139.000000	994.000000	4.000000	9.000000	3.000000	8303.0
75%	698.000000	39957.000000	69.000000	806.000000	148.000000	476.000000	1847.000000	6.000000	13.000000	5.000000	13513.0
max	1952.000000	130100.000000	112.000000	2338.000000	1907.000000	1688.000000	5284.000000	18.000000	45.000000	31.000000	49829.0

1. Exploring Five-number summary of variables.
2. Imputing Missing Values:
   A) For Comorbidities Measure: Default to 0
   B) For Multivalued Categorical: Default to ‘Others’.
   C) For Binary Categories: Random(0,1)
   D) For Continuous: Mean

Identifying good and bad visits

GOOD VISITS

LOS	READMIT	VisitLink
3	0	1671
6	0	1671
2	0	1671
2	0	1671
7	0	2577
7	0	699
4	0	699
4	0	191943
6	0	191943

READMISSION VISITS

INDEX

LOS	READMIT	VisitLink
2	0	753922
8	1	753922
5	1	753922
5	1	753922

LOS	READMIT	VisitLink
6	0	191943
1	1	191943
18	1	191943
4	0	191943
6	0	191943

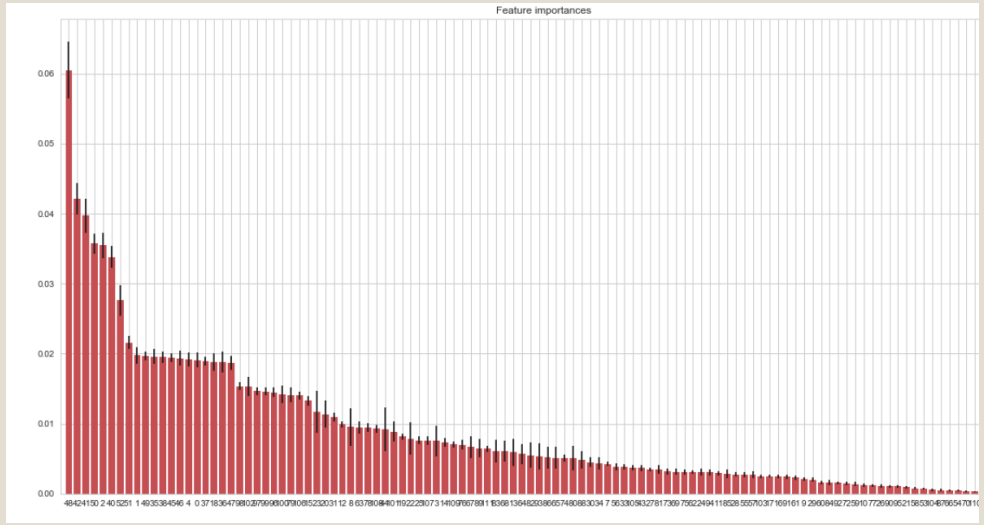
BAD VISITS

LOS	READMIT	VisitLink
4	1	4688420
1	0	4688420
2	0	4688420

LOS	READMIT	VisitLink
14	1	12707
6	1	16131

Feature selection

- EXTRA TREES CLASSIFIER
- RECURSIVE FEATURE ELIMINATION
- STEPWISE AIC

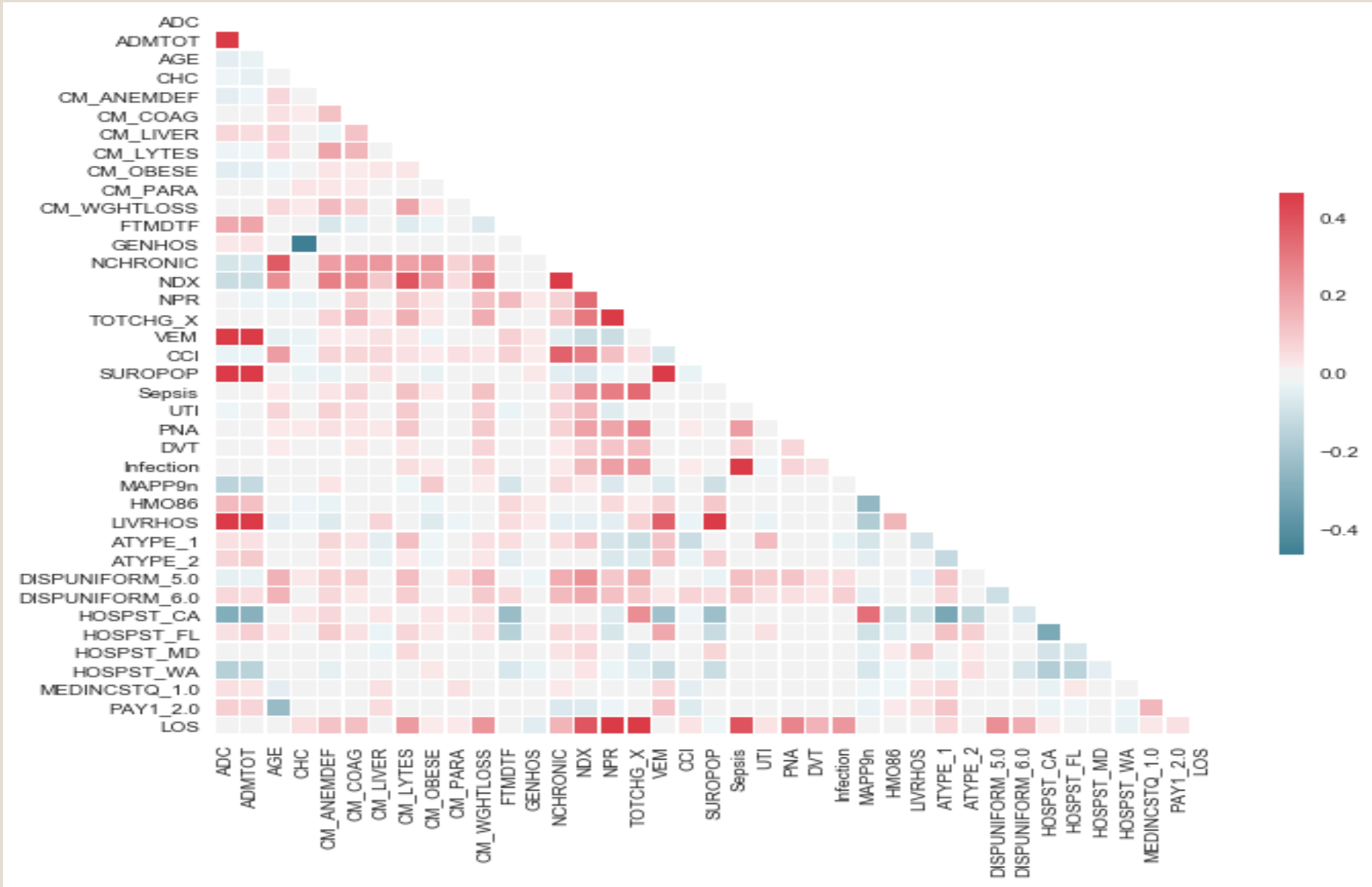



```
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
LOS ~ ADC + ADMTOT + AGE + AWEEKEND + BDTOT + CHC + CM_AIDS +
    CM_ALCOHOL + CM_ANEMDEF + CM_ARTH + CM_BLDLOSS + CM_CHF +
    CM_CHRNLUNG + CM_COAG + CM_DEPRESS + CM_DM + CM_DMCX + CM_DRUG +
    CM_HTN_C + CM_HYPOTHY + CM_LIVER + CM_LYMPH + CM_LYTES +
    CM_METS + CM_NEURO + CM_OBESE + CM_PARA + CM_PERIVASC + CM_PSYCH +
    CM_PULMCIRC + CM_RENLFAIL + CM_TUMOR + CM_ULCER + CM_VALVE +
    CM_WGHTLOSS + FEMALE + FTMDTF + FTRES + FTRNTF + GENHOS +
    NCHRONIC + NDX + NPR + PPO86 + SUROPIP + SUROPOP + SUROPTOT +
    TOTCHG_X + VEM + CCI + PE + MI + Sepsis + UTI + PNA + DVT +
    Infection + Cardiac + Surgery_Complications + MAPP1n + MAPP2n +
    MAPP3n + MAPP5n + MAPP6n + MAPP7n + MAPP8n + MAPP9n + MAPP10n +
    MAPP11n + MAPP12n + MAPP13n + MAPP16n + HMO86 + LIVRHOS +
    MSICHOS + TETOT + MSICBD + ATYPE_1 + ATYPE_2 + ATYPE_3 +
    DISPUNIFORM_1.0 + DISPUNIFORM_2.0 + DISPUNIFORM_5.0 + DISPUNIFORM_6.0 +
    DISPUNIFORM_7.0 + HOSPST_CA + HOSPST_FL + HOSPST_IA + HOSPST_MA +
    HOSPST_MD + HOSPST_NY + HOSPST_WA + HOSPST_WI + MEDINCSTQ_1.0 +
    MEDINCSTQ_2.0 + MEDINCSTQ_3.0 + MEDINCSTQ_4.0 + PAY1_1.0 +
    PAY1_2.0 + PAY1_3.0 + PAY1_4.0 + PAY1_5.0 + PAY1_6.0 + RACE_1.0 +
    RACE_2.0 + RACE_3.0 + RACE_4.0 + RACE_5.0 + RACE_6.0

Final Model:
LOS ~ ADC + ADMTOT + AGE + AWEEKEND + BDTOT + CHC + CM_AIDS +
    CM_ANEMDEF + CM_ARTH + CM_BLDLOSS + CM_CHF + CM_CHRNLUNG +
    CM_COAG + CM_DEPRESS + CM_DRUG + CM_HTN_C + CM_HYPOTHY +
    CM_LIVER + CM_LYMPH + CM_LYTES + CM_METS + CM_NEURO + CM_OBESE +
    CM_PARA + CM_PULMCIRC + CM_RENLFAIL + CM_TUMOR + CM_WGHTLOSS +
    FTMDTF + FTRES + FTRNTF + GENHOS + NCHRONIC + NDX + NPR +
    PPO86 + SUROPIP + SUROPOP + SUROPTOT + TOTCHG_X + VEM + CCI +
    PE + MI + Sepsis + UTI + PNA + DVT + Infection + MAPP1n +
    MAPP2n + MAPP3n + MAPP5n + MAPP6n + MAPP7n + MAPP8n + MAPP9n +
    MAPP10n + MAPP11n + MAPP13n + HMO86 + LIVRHOS + TETOT + MSICBD +
    ATYPE_1 + ATYPE_2 + DISPUNIFORM_2.0 + DISPUNIFORM_5.0 + DISPUNIFORM_6.0 +
    HOSPST_CA + HOSPST_FL + HOSPST_MD + HOSPST_WA + MEDINCSTQ_1.0 +
    MEDINCSTQ_2.0 + MEDINCSTQ_3.0 + PAY1_2.0 + PAY1_3.0 + RACE_1.0 +
    RACE_3.0
```

`feature_imp_rfe`

	importance	var
0	68	ADC
1	79	ADMTOT
2	62	AGE
3	67	AWEEKEND
4	73	BDTOT
5	1	CHC
6	1	CM_AIDS
7	45	CM_ALCOHOL
8	1	CM_ANEMDEF
9	3	CM_ARTH
10	29	CM_BLDLOSS
11	1	CM_CHF
12	49	CM_CHRNLUNG
13	5	CM_COAG

CORRELATION MATRIX



Data modeling

Evaluating Model Performance

Ordinary Least Squares

	Fold1	Fold2	Fold3	Fold4	Fold5	**Average**
Adjusted R sq.	80.7	80.8	80.8	81.5	80.9	**80.94**
RMSE	4.27	4.006	4.43	4.17	4.23	**4.2212**

Final OLS Model - TEST MODEL

RMSE	4.19923
MSE	17.63353
Abs Mean	2.225504

Multi Linear Regression

	Fold1	Fold2	Fold3	Fold4	Fold5	Fold6	Fold7	Fold8	Fold9	Fold10	**Average**
Adjusted R sq.	0.698918	0.700705	0.653858	0.632295	0.665496	0.635726	0.622774	0.58199293	0.47077	0.65798	**0.632**

Random Forest

RMSE	5.809707	R sq Train	0.4886
MSE	33.75269	R sq Test	0.5518
Absolute Mean	3.432953		

Support Vector Regression

RMSE	5.809707	R sq Train	0.285622
MSE	33.75269	R sq Test	0.298241
Absolute Mean	3.432953		

CALCULATION OF COST SAVED

- Cost Saved per VisitLink =

CHG_Per_Visit * (Total_LOS (Including READMITS following an Index Visit) – Ideal_LOS (Predicted))

- CHG_Per_Visit = $\frac{\text{Total\_CHG}}{\text{Total\_LOS}}$