Capstone Project Final Presentation

New York State Hospital Inpatient Discharge

DSE Online July'22 - Group 6

Mentor: Vidhya K.

Team Leader: Shailja Barsaiyan

Team Members:

Aditya Aryan
Mahima Bhardwaj
Morvi Kohad
Shailja Barsaiyan
Shrika Garg

greatlearning

Problem Definition & Statement

- **Healthcare industry in New York (NY) -** facing difficulty in ensuring that every community across the state has equal access to healthcare.
- Examine hospital inpatient discharge data from diverse communities in New York, to **identify regions where additional healthcare coverage is needed** for specific diagnoses or procedure codes.
- Healthcare providers can **allocate their resources** more efficiently and enhance healthcare outcomes.

Hospital Service Area	Hospital County	Operating Certificate Number	Permanent Facility Id	Facility Name	Age Group	Zip Code - 3 digits	Gender	Race	Ethnicity
Long Island	Nassau	2950001.0	527.0	Mount Sinai South Nassau	70 or Older	115	M	White	Not Span/Hispanic
Long Island	Suffolk	5153000.0	913.0	Huntington Hospital	70 or Older	117	F	Black/African American	Not Span/Hispanic
New York City	Richmond	7004003.0	1737.0	Staten Island University Hospital Prince's Bay	50 to 69	103	M	White	Spanish/Hispanic

Problem Statement 1

Predicting **Length of Stay** through a Machine Learning Supervised Algorithm.

Problem Statement 2

To anticipate the likelihood of **Patient Mortality**.

Problem Statement 3

Analyse demographics to **identify specific diagnoses.**

Dataset

- Hospital inpatient discharge data by Dept. of Health of New York State publicly available at: https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/tg3i-cinn
- The dataset consists of **2101588 rows** & **33 columns**. For simplification, considered **2,00,000 rows** from our dataset.

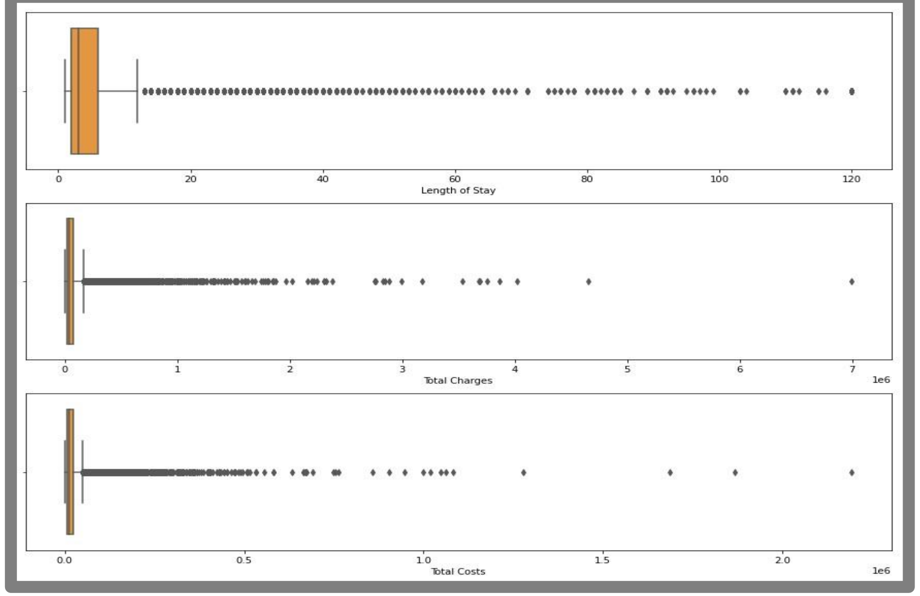
```
RangeIndex: 210158 entries, 0 to 210157
Data columns (total 33 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   Hospital Service Area               209092 non-null  object
 1   Hospital County                     209092 non-null  object
 2   Operating Certificate Number        208954 non-null  float64
 3   Permanent Facility Id               209092 non-null  float64
 4   Facility Name                       210158 non-null  object
 5   Age Group                           210158 non-null  object
 6   Zip Code - 3 digits                 205617 non-null  object
 7   Gender                              210158 non-null  object
 8   Race                                210158 non-null  object
 9   Ethnicity                           210158 non-null  object
 10  Length of Stay                      210158 non-null  int32
 11  Type of Admission                   210158 non-null  object
 12  Patient Disposition                 210158 non-null  object
 13  Discharge Year                      210158 non-null  int64
 14  CCSR Diagnosis Code                 210012 non-null  object
 15  CCSR Diagnosis Description          210012 non-null  object
 16  CCSR Procedure Code                 152647 non-null  object
 17  CCSR Procedure Description          152647 non-null  object
 18  APR DRG Code                        210158 non-null  int64
 19  APR DRG Description                 210158 non-null  object
 20  APR MDC Code                        210158 non-null  int64
 21  APR MDC Description                 210158 non-null  object
 22  APR Severity of Illness Code        210158 non-null  int64
 23  APR Severity of Illness Description 209909 non-null  object
 24  APR Risk of Mortality               209909 non-null  object
 25  APR Medical Surgical Description    210158 non-null  object
 26  Payment Typology 1                  210158 non-null  object
 27  Payment Typology 2                  102735 non-null  object
 28  Payment Typology 3                  33282 non-null   object
 29  Birth Weight                        20806 non-null   object
 30  Emergency Department Indicator      210158 non-null  object
 31  Total Charges                       210158 non-null  float64
 32  Total Costs                         210158 non-null  float64
dtypes: float64(4), int32(1), int64(4), object(24)
```

Exploratory Data Analysis

- Found and handled **missing data**
- 757 duplicate rows were dropped.
- Outliers – Box-Cox Transformation
- Removed **redundant columns.**

```
(health_data.isnull().mean()*100).sort_values(ascending=False)
```

```
Birth Weight                                90.187052
Payment Typology 3                          84.163254
Payment Typology 2                          51.140906
CCSR Procedure Code                         27.293285
CCSR Procedure Description                  27.293285
Zip Code - 3 digits                          2.133174
Operating Certificate Number                 0.581053
Hospital County                              0.514515
Hospital Service Area                        0.514515
Permanent Facility Id                        0.514515
APR Risk of Mortality                        0.117989
APR Severity of Illness Description          0.117989
CCSR Diagnosis Code                          0.074795
CCSR Diagnosis Description                   0.074795
```



greatlearning

Challenges

- Gender: 'U' -> 'F'.
- Type of Admission: 'Not available' -> 'Emergency'.
- Ethnicity: 'Unknown' -> 'Spanish/Hispanic'.
- Length of Stay: '120 +' -> '120'.
- Zip code- 3 digits: 'OOS' -> '999'.

```
health_data.Gender= health_data['Gender'].str.replace('U','F')
```

```
health_data['Type of Admission'] = health_data['Type of Admission'].str.replace('Not Available','Emergency')
```

```
health_data['Length of Stay'].unique()
```

```
array(['4', '8', '1', '16', '2', '19', '14', '11', '13', '20', '3', '9',
       '6', '5', '12', '17', '15', '22', '34', '7', '25', '23', '21',
       '10', '24', '30', '35', '78', '18', '33', '26', '42', '41', '81',
       '31', '48', '28', '29', '38', '56', '49', '103', '27', '91', '59',
       '120 +', '53', '39', '52', '40', '46', '51', '32', '60', '45',
       '58', '44', '43', '66', '36', '50', '85', '63', '102', '68', '61',
       '84', '89', '54', '47', '37', '76', '92', '64', '71', '95', '62',
       '83', '74', '67', '111', '99', '80', '55', '75', '96', '82', '87',
       '98', '104', '77', '110', '69', '112', '97', '93', '115', '116',
       '57', '114', '72', '65', '79', '70', '86', '118', '94', '105',
       '117', '73', '109', '107', '108', '101', '106', '90', '119', '113',
       '100', '88'], dtype=object)
```

	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	PCA7	PCA8	PCA9
0	0.994469	-1.832399	-0.558417	0.702948	0.140031	-1.423593	-0.862880	-0.967540	0.744109
1	0.986631	-1.054495	-1.657258	-0.380080	1.021009	-0.411818	-0.916541	0.272960	-0.069503
2	-0.714474	1.208255	0.564185	-2.231641	-0.065607	0.787736	-2.306795	-1.721018	1.495781
3	-0.330712	-0.880742	1.419413	-3.961436	0.283889	-0.150569	1.488955	-2.304413	2.817339
4	1.026106	0.182703	-1.012931	-0.088671	1.076888	0.347429	1.875579	0.673722	-0.800719

```
health_data[health_data['Zip Code - 3 digits'] == 'OOS'].count()[0]
```

```
17731
```

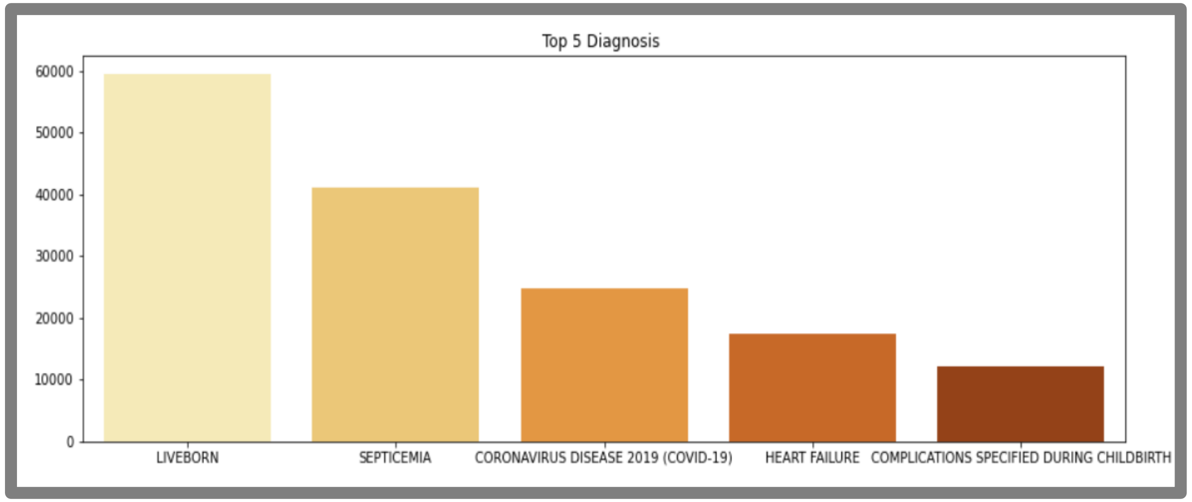
```
health_data['Zip Code - 3 digits'] = health_data['Zip Code - 3 digits'].apply(lambda x: str(x).replace('OOS','999'))
health_data['Zip Code - 3 digits'] = health_data['Zip Code - 3 digits'].astype(np.number)
```

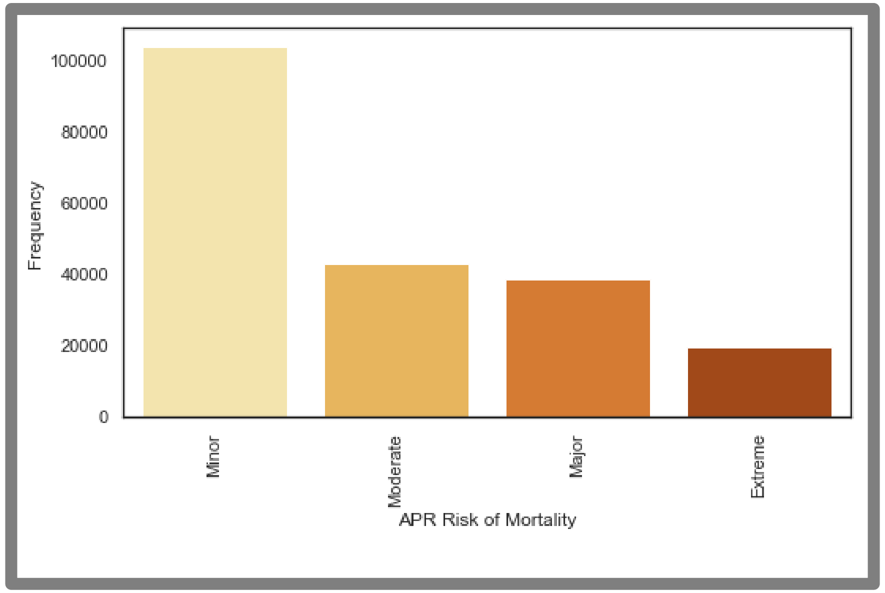
```
Before OverSampling, counts of label '0': 72723
Before OverSampling, counts of label '1': 30114
Before OverSampling, counts of label '2': 27083
Before OverSampling, counts of label '3': 13686


After OverSampling, counts of label '0': 72723
After OverSampling, counts of label '1': 72723
After OverSampling, counts of label '2': 72723
After OverSampling, counts of label '3': 72723
```

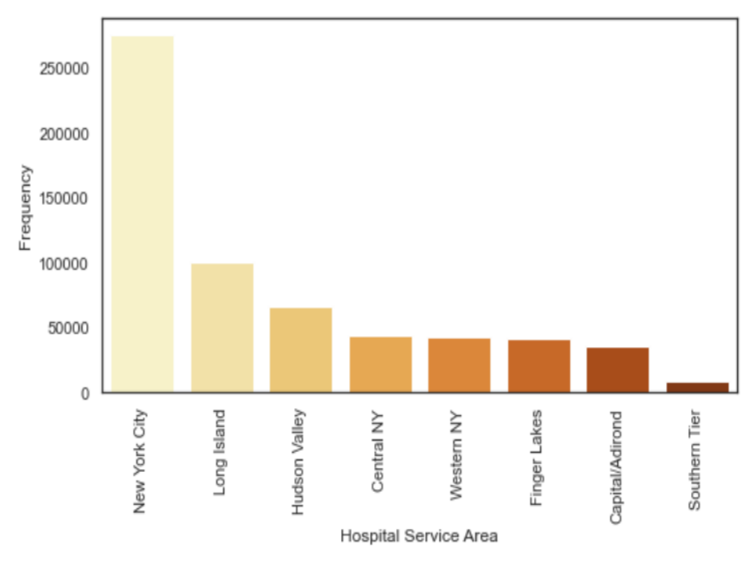
- Performed PCA for dimensionality reduction - 70% variation
- Class balancing through SMOTE

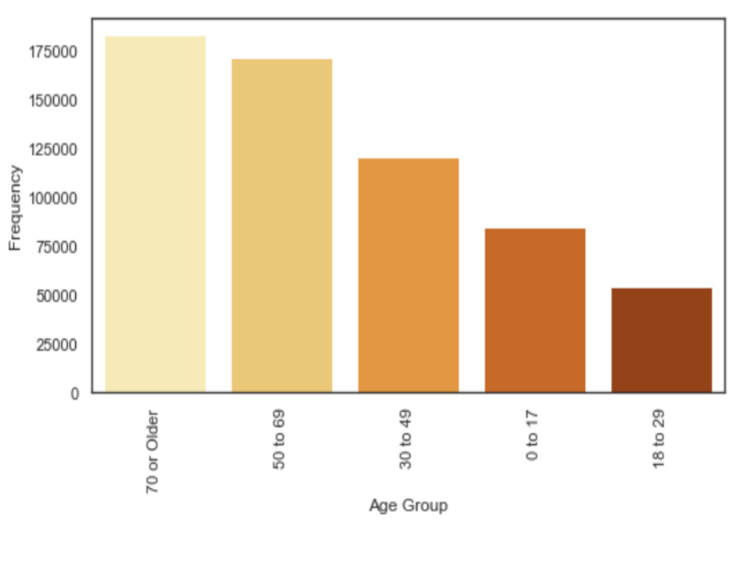
Univariate Analysis

- Most people were diagnosed with Liveborn.
- Most of the patients are least likely to die.
- No of patients are much more in New York city.
- Most people belong to 70-older age group.



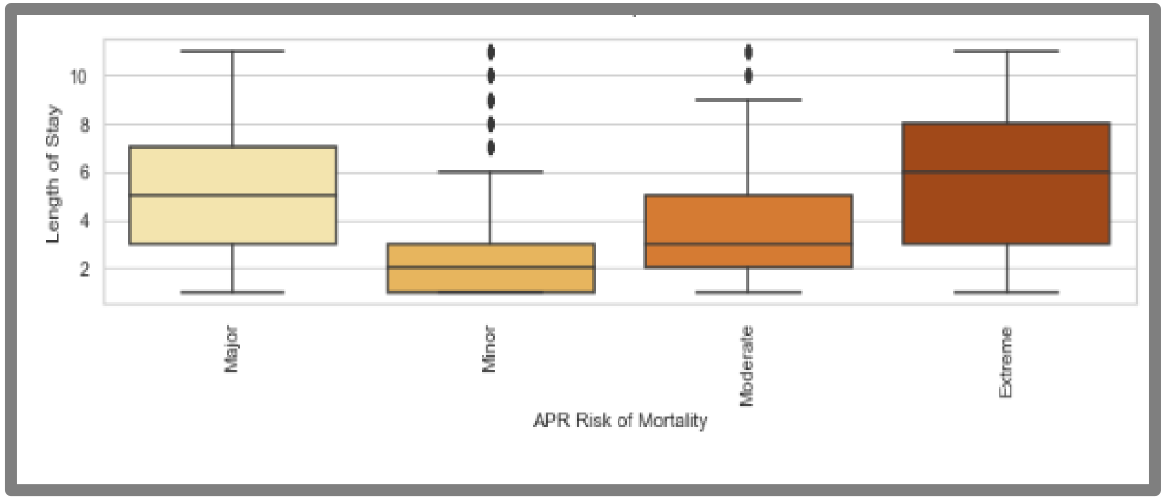


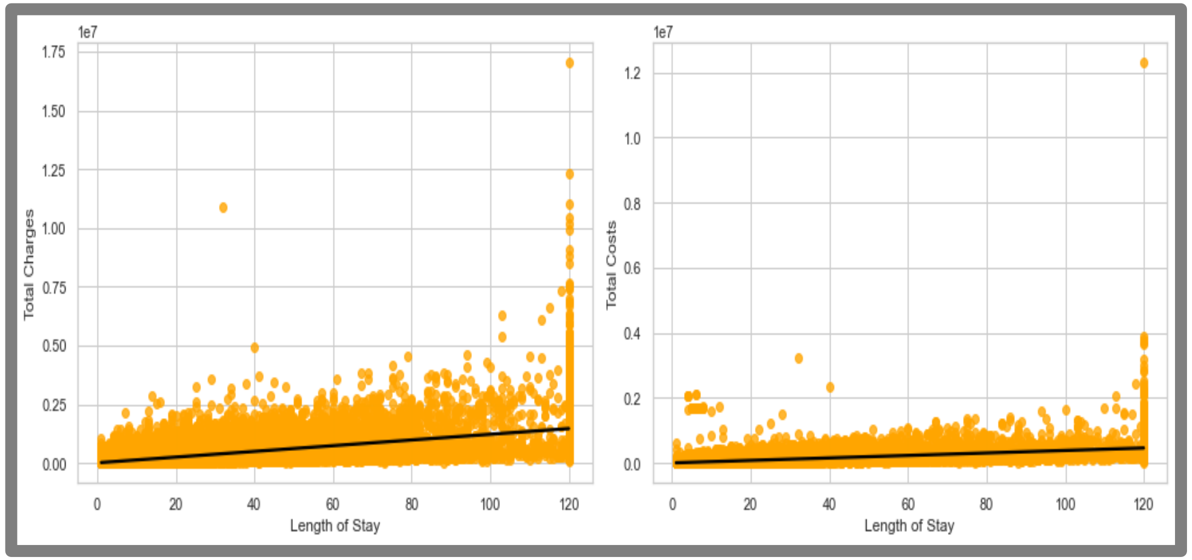


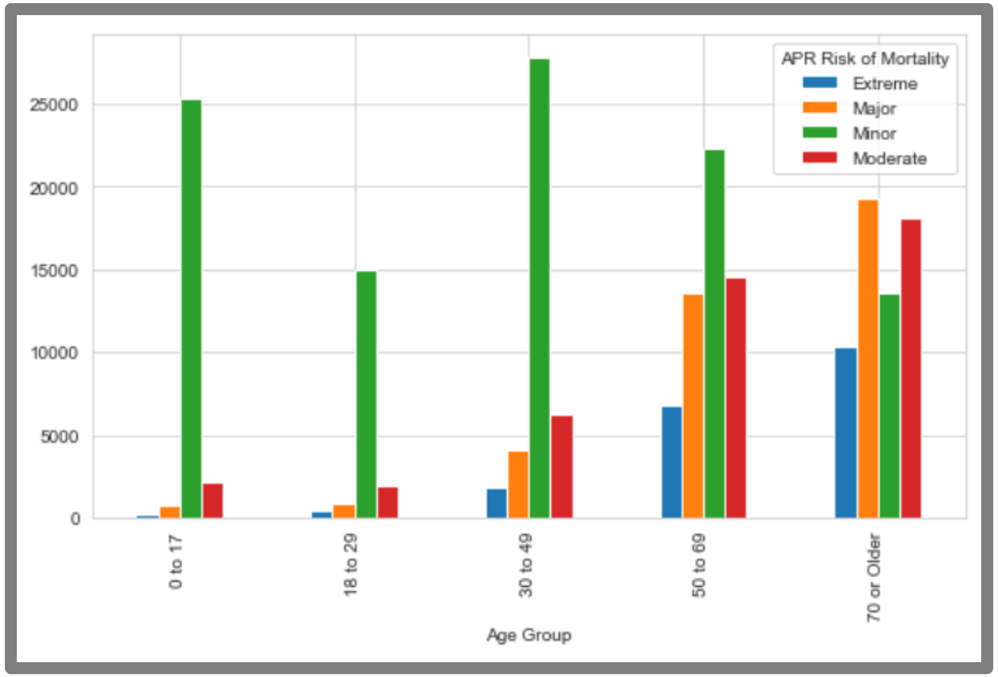


Bivariate Analysis

- People with lowest risk of mortality have shortest stays.
- Total charges & costs have positive relation with LOS.
- Risk of Mortality increases with respect to age.



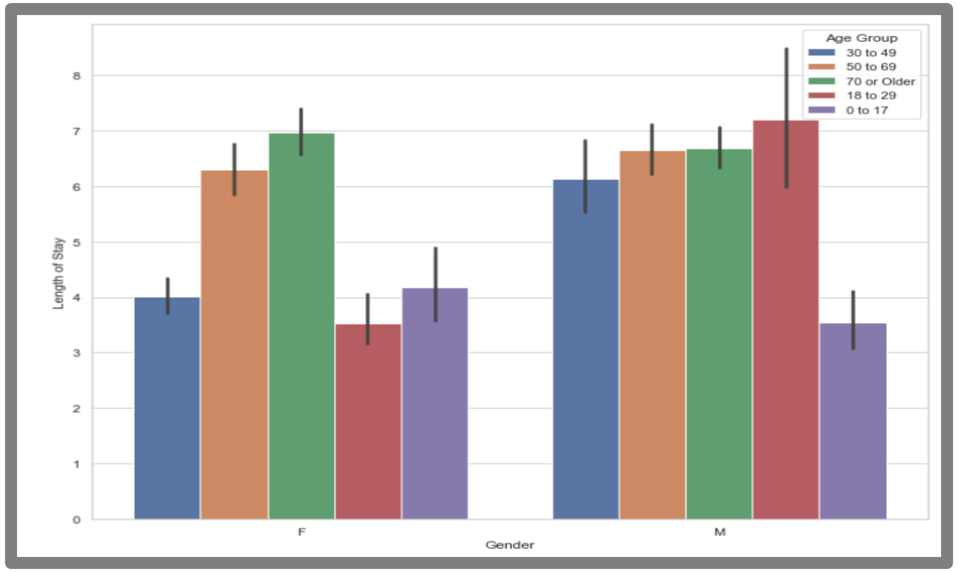


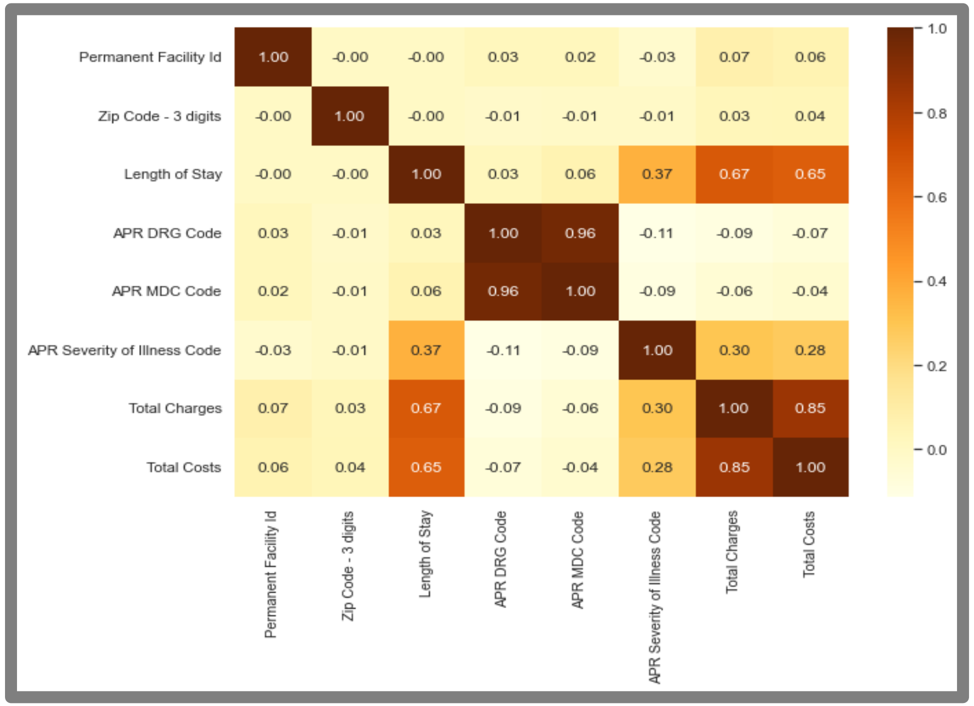


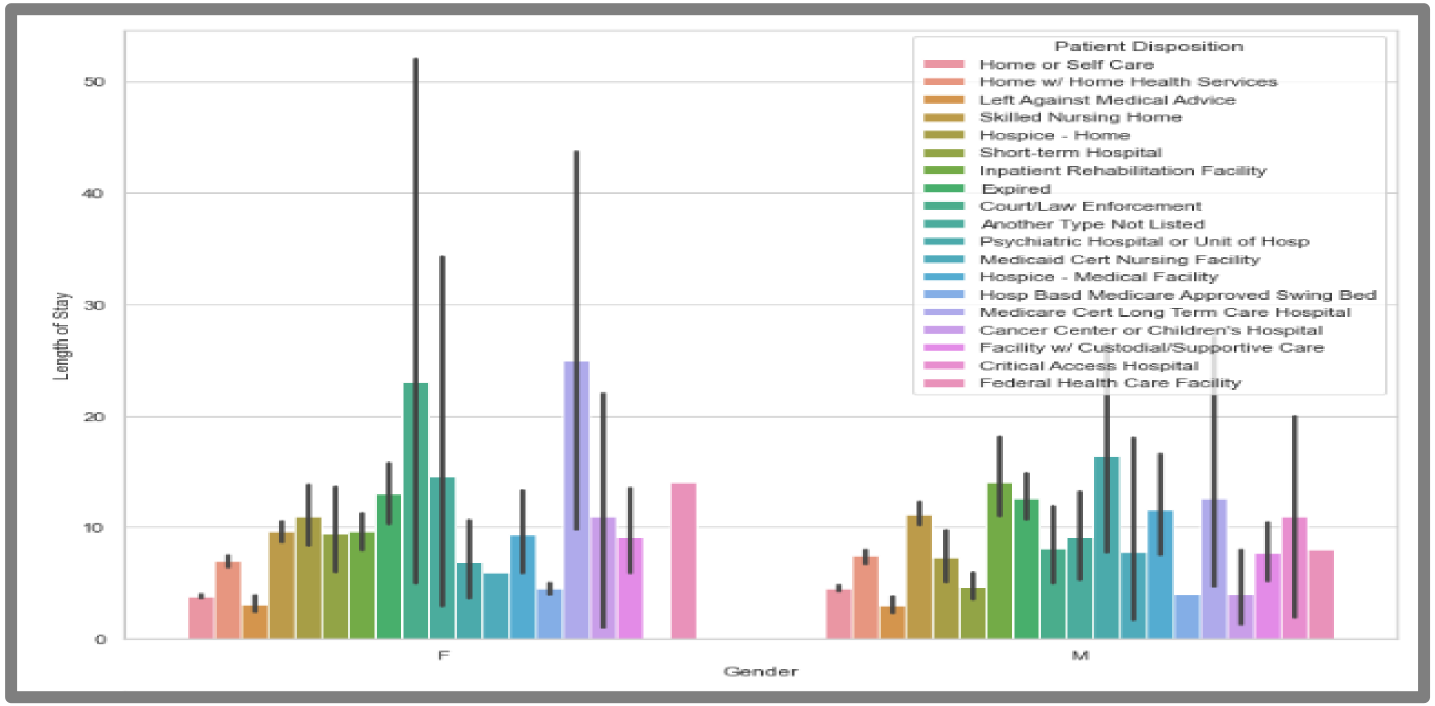
greatlearning

Multivariate Analysis

- Males above 17 are staying more than females.
- Cancer Center or Children Hospital Females stayed longer than men on the same prescription.
- Total charges and costs have positive correlation with length of stay.







greatlearning

Regression

Predicting Length of Stay

XGBoost

- ✓ Faster than other algorithms
- ✓ Top performing for large datasets

- × High computation
- × Loss of explanatory power
- × Prone to overfitting

	Model	R2-score	MSE	RMSE	MAPE
8	XGBoost	8.186505e-01	1.215737e+01	3.486742e+00	4.257394e-01
6	Random Forest	6.470432e-01	2.366165e+01	4.864324e+00	6.569270e-01
5	Decision Tree	6.224458e-01	2.531062e+01	5.030966e+00	6.808814e-01
2	Ridge Regression	4.719058e-01	3.540257e+01	5.950006e+00	1.097963e+00
0	Linear Regreesion	4.719058e-01	3.540257e+01	5.950006e+00	1.097971e+00
1	Lasso regression	4.621198e-01	3.605861e+01	6.004882e+00	1.143252e+00
4	KNeighbors	3.842834e-01	4.127663e+01	6.424689e+00	7.520471e-01
7	Adaboost	3.817232e-01	4.144826e+01	6.438032e+00	1.881566e+00
3	SGDRegressor	-8.924664e+26	5.982949e+28	2.446007e+14	7.957311e+13

```
building_model(XGBRegressor(max_depth = 7,learning_rate=0.03,n_estimators= 150),X_train,X_test,y_train,y_test)

R2 score on test data:  0.845361880942902
R2 score on train data:  0.889489995935017
RMSE Traning: 2.7437
RMSE Testing: 3.2197
```

Model Building
- Encoding and Scaling
- Performed 9 Algorithms
- Highest R2 Model - XGBoost

GridSearchCV
- Hyperparameter Tuning – Found out best parameters.

Tweaking the Parameters
- Tweaked the best parameters to reduce overfitting.

Final Model
- Chose the final model with an R2 of 0.85.

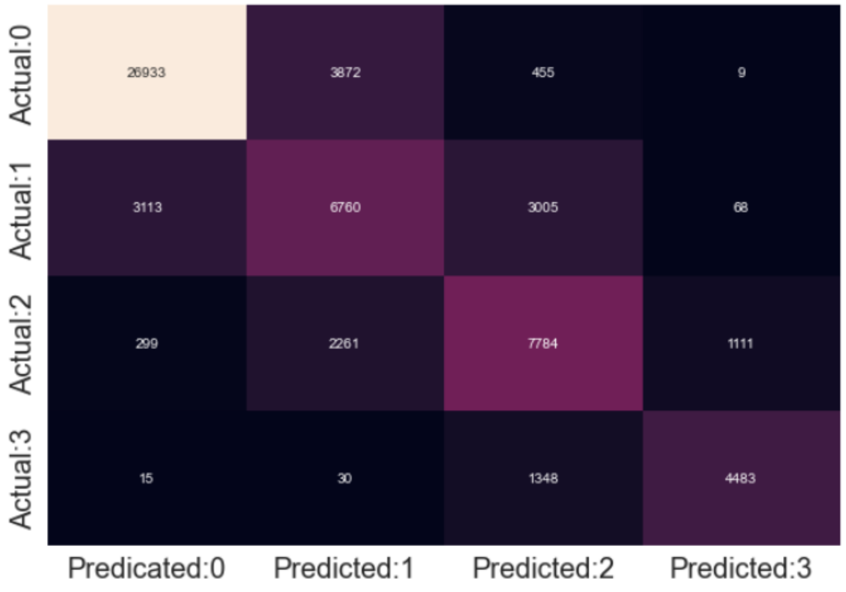
greatlearning

Classification

Anticipating Risk of Mortality

	precision	recall	f1-score	support
0	0.89	0.86	0.87	31269
1	0.52	0.52	0.52	12946
2	0.62	0.68	0.65	11455
3	0.79	0.76	0.78	5876
accuracy			0.75	61546
macro avg	0.70	0.71	0.71	61546
weighted avg	0.75	0.75	0.75	61546

	Model	Accuracy	Precision	Recall	F1-score	Cohen-Kappa
6	XGBoost	0.749147	0.706064	0.704929	0.705056	0.617180
4	Random Forest	0.734719	0.701913	0.708635	0.703087	0.604168
3	Decision Tree	0.730852	0.689986	0.694400	0.691491	0.594090
0	Logistic Regreesion	0.714977	0.683174	0.695089	0.686625	0.576730
1	SGD Classifier	0.667192	0.560562	0.594824	0.551555	0.480408
2	Naive Bayes	0.658532	0.629032	0.643087	0.630866	0.498932
5	Adaboost	0.655721	0.620615	0.629731	0.624279	0.482554




```
Train accuracy = 0.7958383180011825
Test accuracy = 0.7467585220810451
Recall = 0.706490969627444
Precision = 0.7047257419365186
F1 score = 0.7051368560764072
Kappa = 0.6153310155288524
```

Model Building

- Class Balancing using SMOTE
- Performed 7 Algorithms
- Highest Accuracy Model - XGBoost

GridSearchCV

- Hyperparameter Tuning – Found out the best parameters.

Tweaking the Parameters

- Tweaked the best parameters to improve precision and recall.

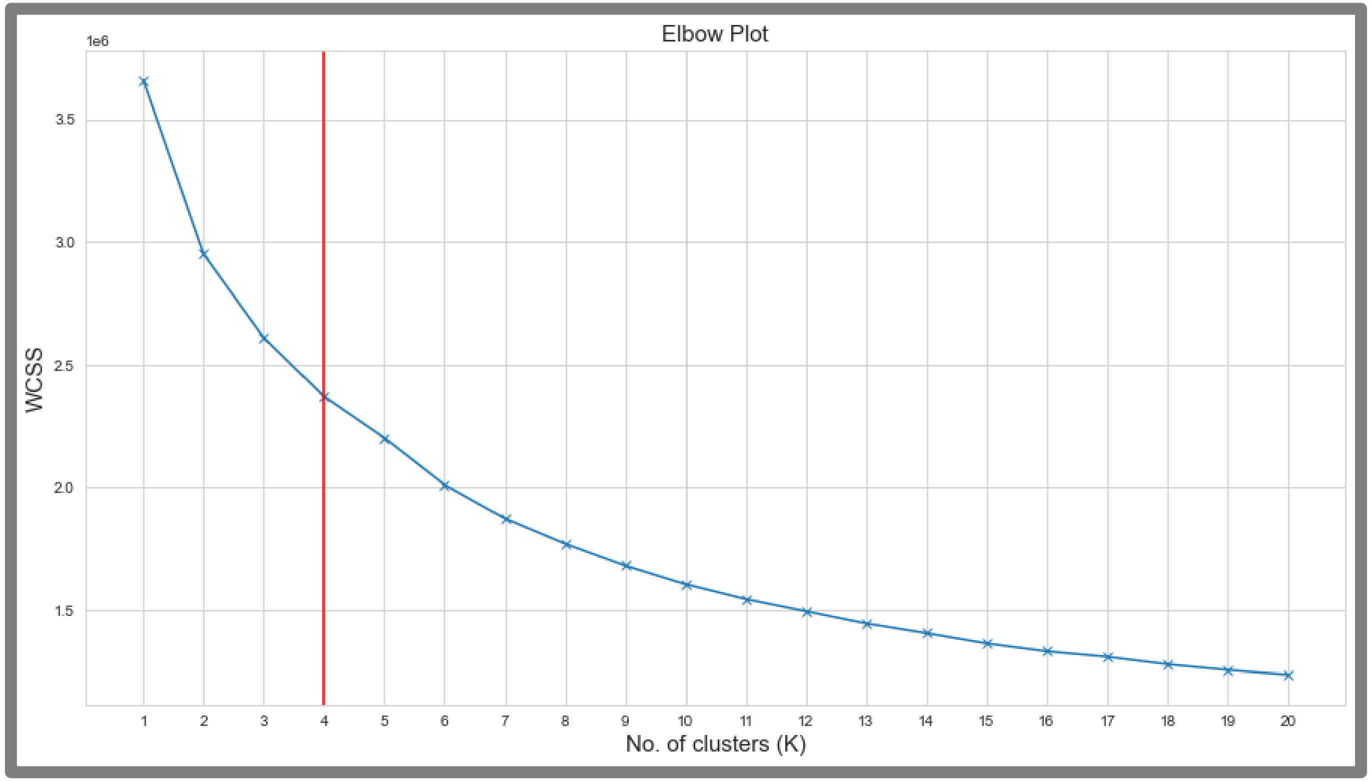
Final Model

- Chose the final model with an accuracy of 0.75.

Clustering

K-Means Clustering

- ✓ Simple
- ✓ Eliminates bias
- ✓ Great for large scale data
- × High computation
- × Sensitive to initial centroids
- × Prone to curse of dimensionality



PCA
- Scaling
- Finding Eigen values and vectors
- Analyzing % variations
- Choosing 9 components

Elbow Plot
- WCSS vs K for up to 20 clusters
- Elbow point at K = 4

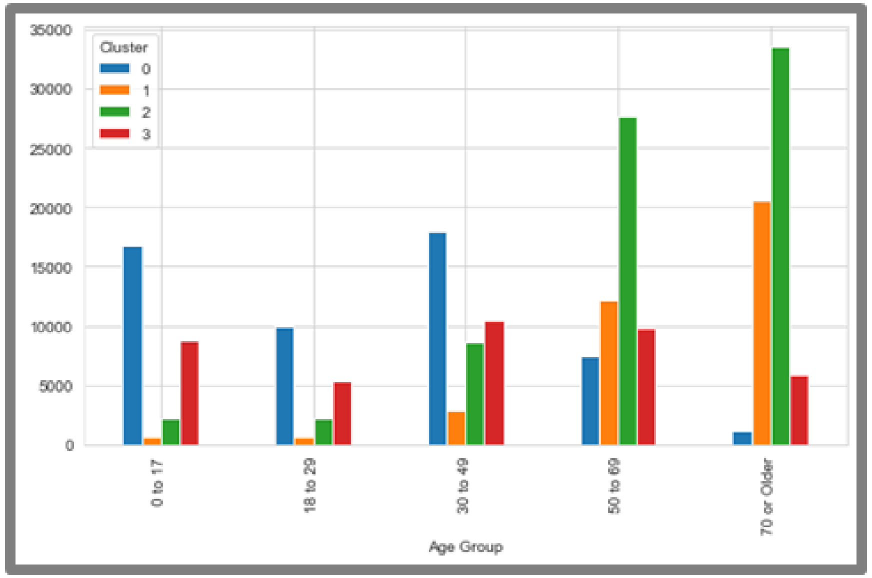
Silhouette Score Analysis
- Finding silhouette score for up to 10 clusters

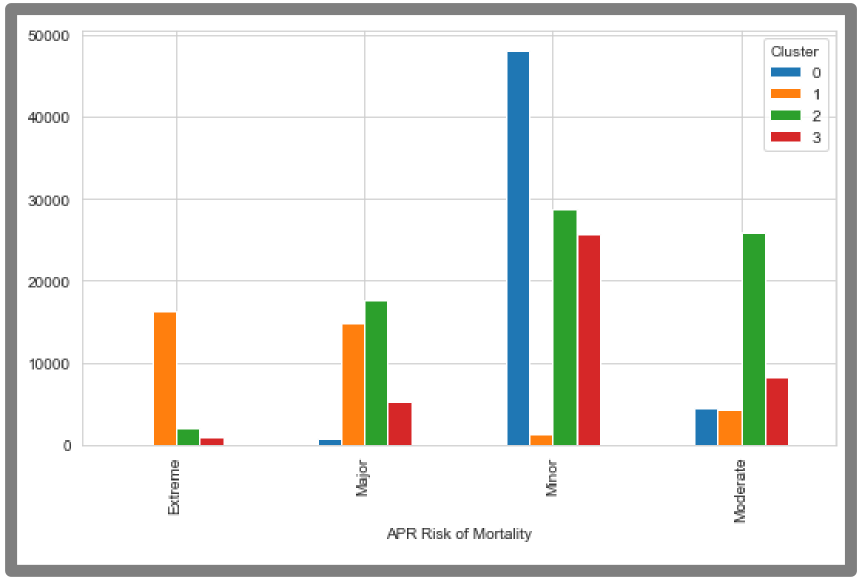
K-Means
- Building model with K = 4
- Adding labels to data

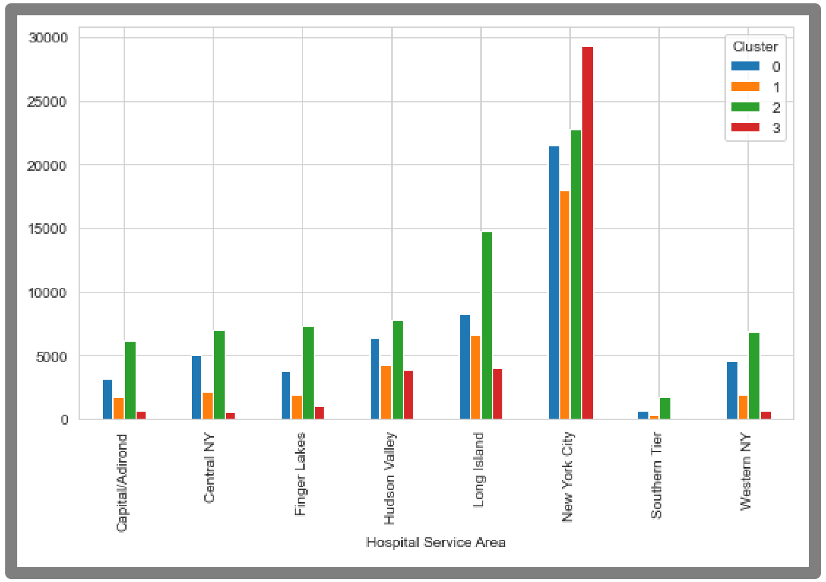
greatlearning

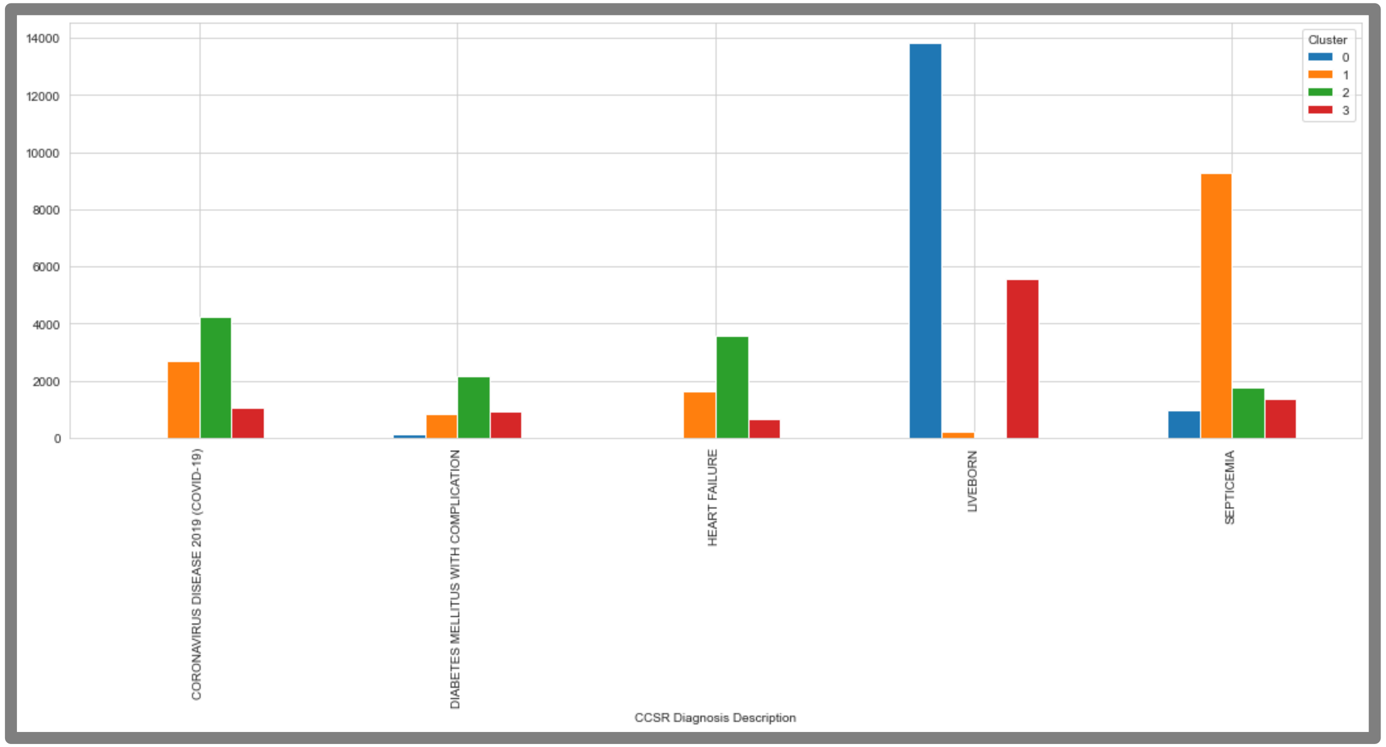
Clustering

- **Cluster 0** belongs to the subgroup that impacts livebirth - Not Spanish/Hispanic ethnicity.
- **Cluster 1** belongs to the subgroup that impacts Septicaemia.
- **Cluster 2** belongs to the subgroup that impacts Covid disease.
- **Cluster 3** belongs to the subgroup that impacts livebirth - Spanish/Hispanic ethnicity.









Inferences

Regression (LOS)

- Children below 18 have shorter stays.
- New-born admissions take 2 days.
- Advance resource allocation for severely ill and high-risk patients is needed.
- Patients admitted in emergencies tend to have longer stays, therefore separate rooms can be reserved for them.
- Our regression model R2 suggests that it explains 84.5% of variation in LOS; with a RMSE of 3.2.

Classification (Risk of Mortality)

- Hospitals should prioritize specialized care for older patients (70 and above) due to their higher mortality risk.
- Patients with 'Medicare' payment typology have higher mortality risks, suggesting the need for improved care coordination.
- Strategies to reduce mortality risks should be prioritized for patients admitted in emergency situations.
- Our classification model classifies 75% of unseen data correctly; with precision and recall of 70% and 71% respectively.

Inferences

Clustering

- Manhattan County and New York City hospitals should collaborate with local health authorities to address higher mortality risks collectively.
- Implement robust management protocols due to the prevalence of septicaemia diagnoses.
- Prioritize patient safety initiatives, including quality improvement programs, infection prevention and control measures.
- Clusters are formed which describe livebirth, Septicemia and COVID diagnoses.

Further Steps

- In the future, we can focus on ensuring the availability and quality of relevant data to train the models effectively.
- Additionally, we can explore different feature engineering techniques to enhance the predictive power of the models.
- Before deploying, we can check our model's performance on a diverse set of data to ensure it generalizes well, test it against edge cases, outliers, and scenarios that may challenge its accuracy.

Thank You!!