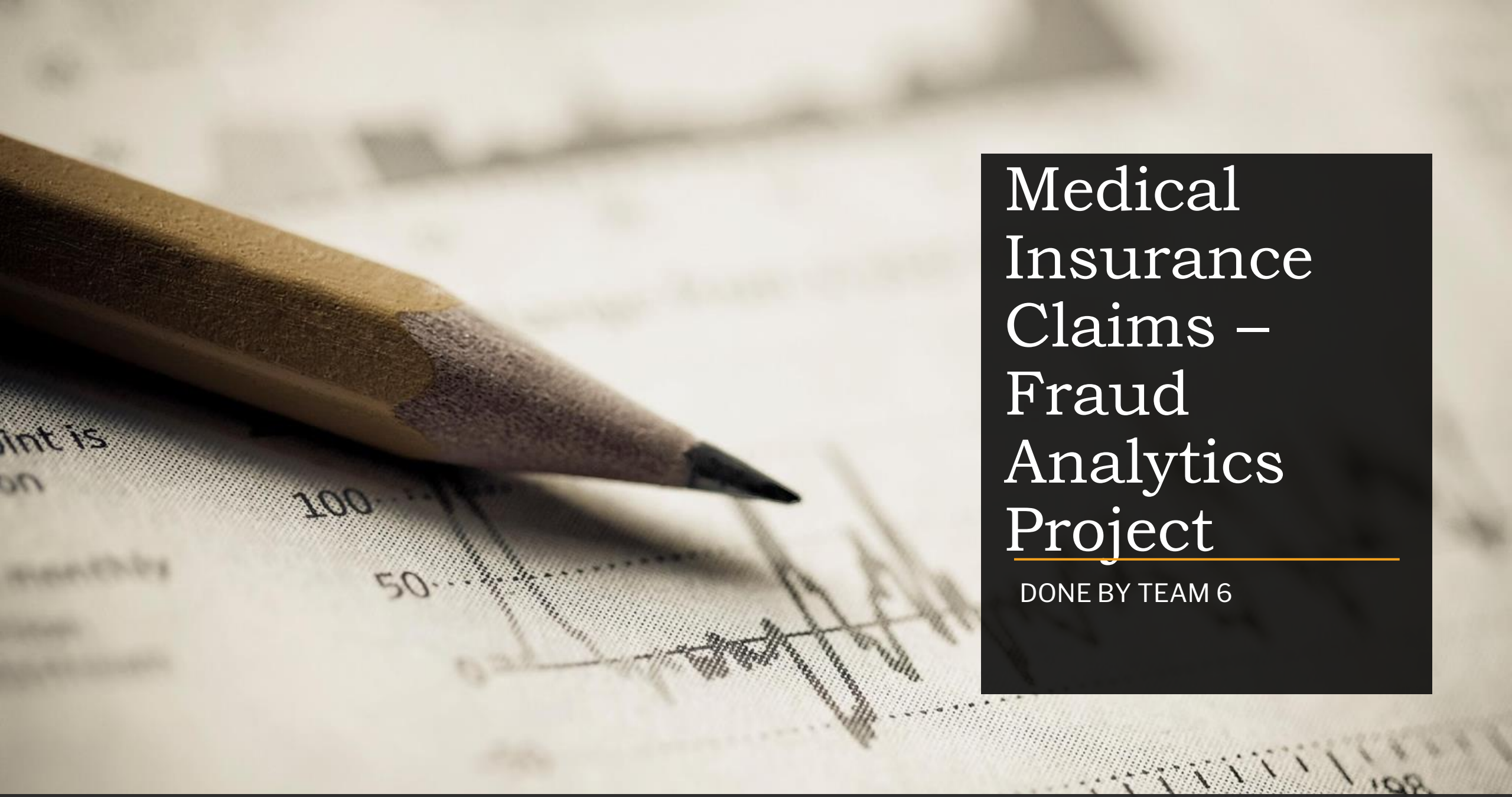
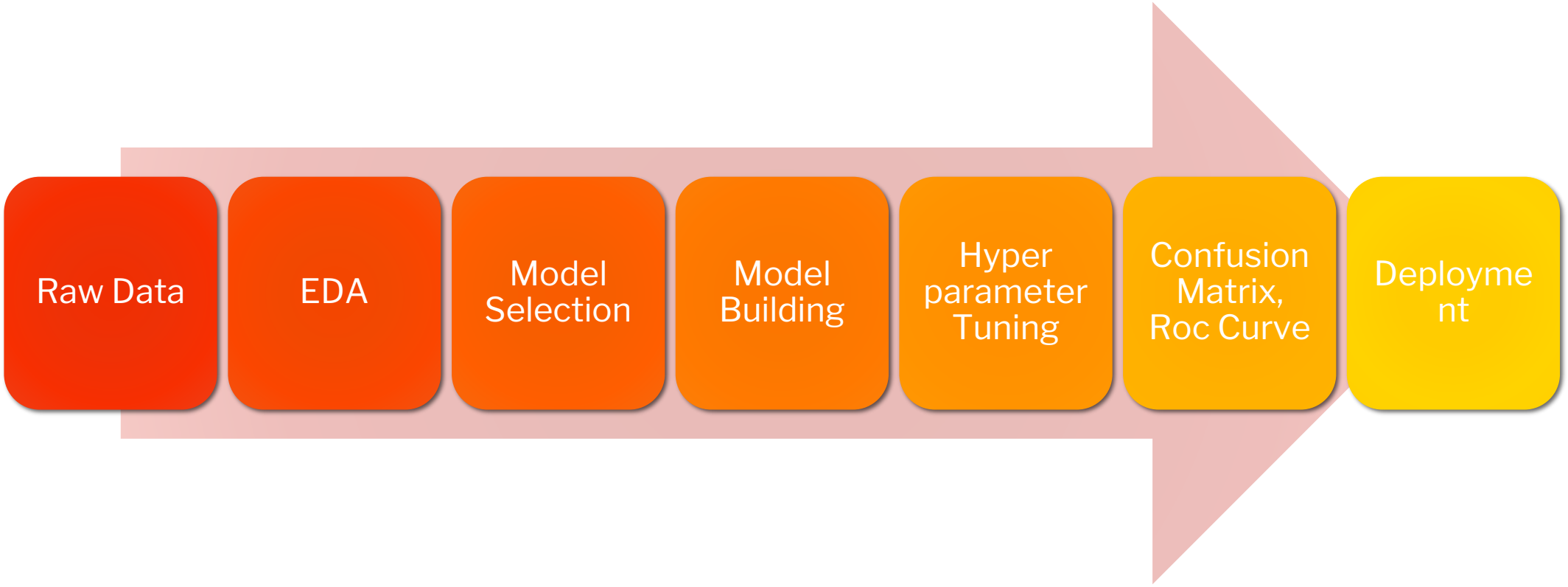


Medical Insurance Claims – Fraud Analytics Project

DONE BY TEAM 6

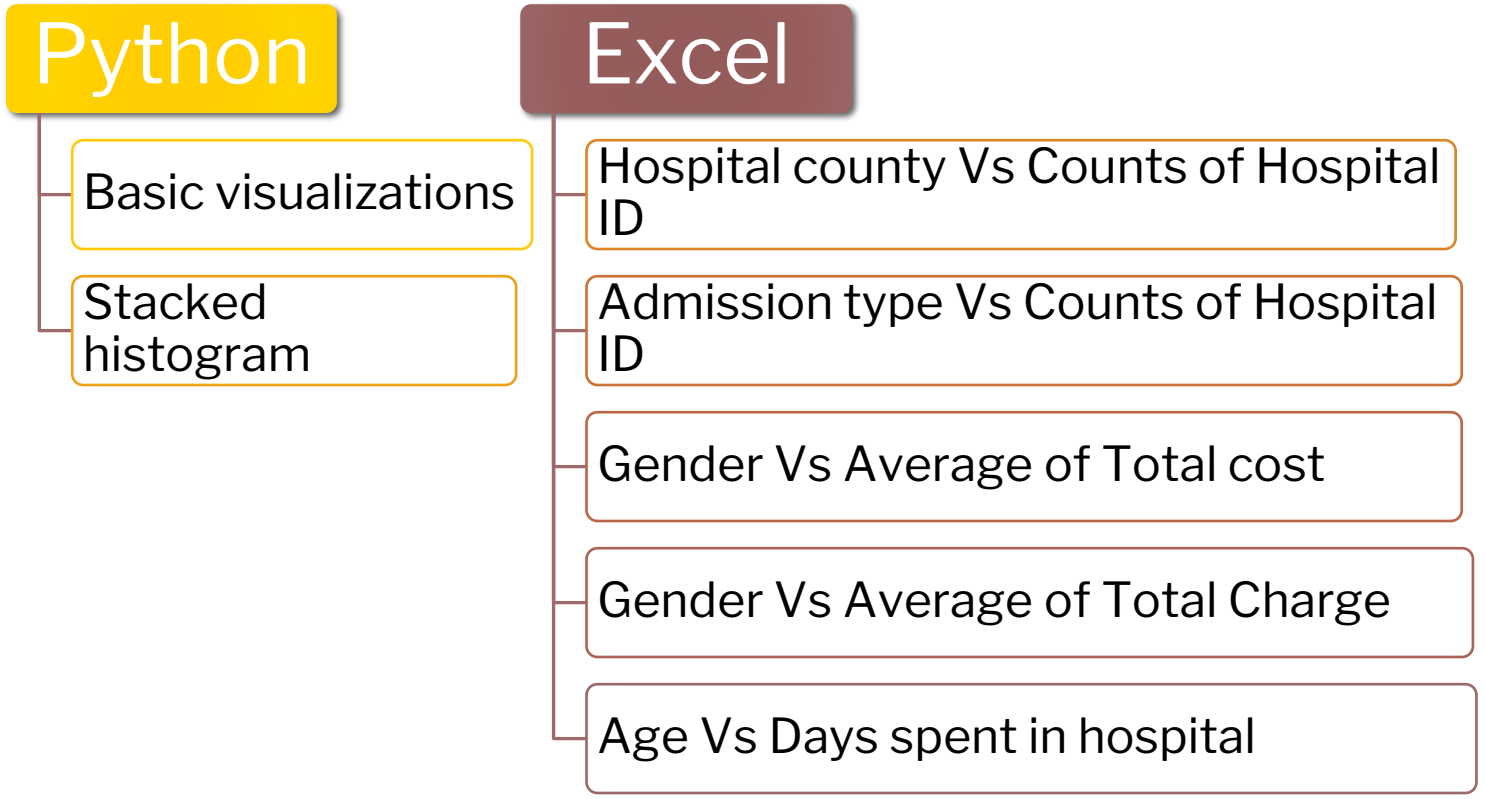
Process Flowchart (Road Map)



Process Flowchart (Road Map)

EDA	MODEL SELECTION	MODEL BUILDING & TUNING	DEPLOY MENT
1. Basic understanding of data. 2. Data cleaning (Missing & duplicate values treatment). 3. Data Visualizations (using Excel & Python) 4. Analyzing interactions & Correlation checking (Input parameter vs input parameter & Input parameter vs output parameter) 5. Label Encoding & Feature selection	1. Approach to Imbalance dataset. (Test & Train) 2. Under sampling (RandomUnderSampler). 3. Over sampling (RandomOverSampler & Smote) 4. Selection of best model based on Mean % score, Bias & Variance (Logistic Regression, DecisionTreeClassifier, XGBClassifier, GaussianNB & RandomForestClassifier) 5. Model Score [mod_score] ('Train Score accuracy', 'Test Score accuracy', 'Recall Score', 'Precision Score', 'F1-Score') (Selecting Top 2 Models by Trial-error on values like test_size, sampling_strategy, criterion & few etc.)	1. Model Building (XGB & RandomForestClassifier[RFC]) 2. Hyper parameter tuning (Selecting best parameters) 2(I). XGB (max_depth; min_child_weight; n_estimators) 2(II). RFC (n_estimators; criterion[gini,entropy];max_features['auto','log2','sqrt'];max_leaf_nodes[10,12,14,15]) 3. Outcome: RFC stands best.	1. Tuned Model score. 2. Confusion Matrix. 3. Roc Curve. 4. Deployment

Visualizations

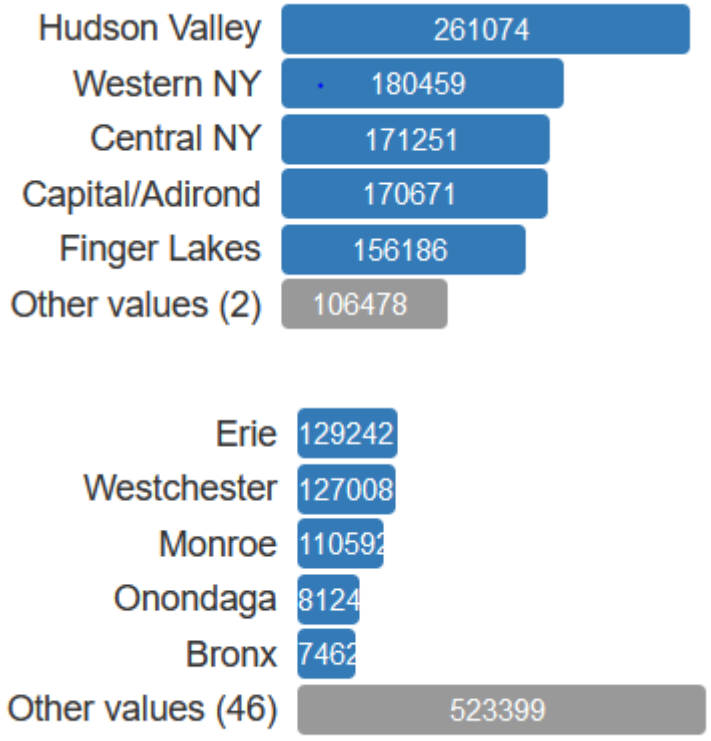


Visualizations (Description)

1. Visualizations using Pandas Profiling (Python)
   - Basic visualizations on attributes
   - Stacked histogram (Input parameters Vs Output parameter [Result])
2. Visualizations by Pivot chart (Excel) {Input parameter Vs Input parameter}

   [Note: Some parameters here are with respect to each Area services]
   - Hospital county Vs Counts of Hospital ID (with respect to Result)
   - Admission type Vs Counts of Hospital ID (with respect to Result & Gender)
   - Gender Vs Average of Total cost (with respect to Result & Mortality Risk)
   - Gender Vs Average of Total Charge (with respect to Result & Mortality Risk)
   - Age Vs Days spent in hospital (with respect to Result & Gender)

Basic visualizations on attributes



Areas of service :

There are 7 areas of services where data on hospital facilities are collected

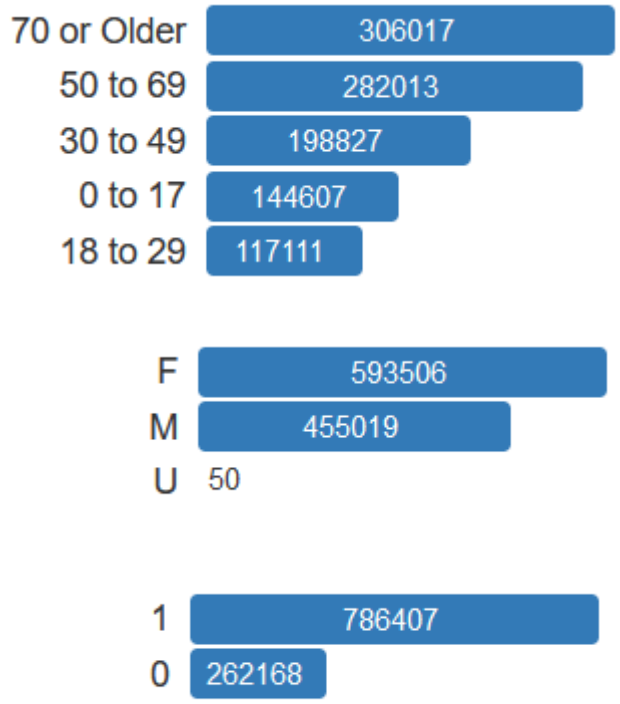
Hospital County :

There are certain counties in each area where the hospital services are provided

Hospital Id :

In each county, there are various types of hospitals with different Id numbers

Basic visualizations on attributes



Age :

There are patients of different age groups starting from 0 to 70 years and older. This includes newborn baby as well

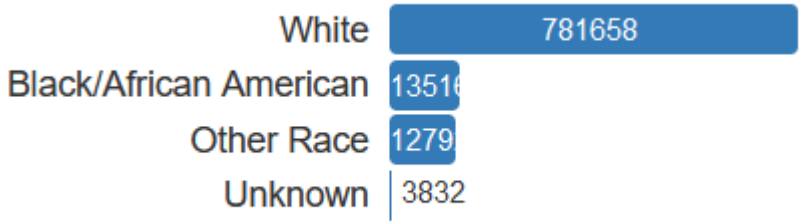
Gender: Male/ Female

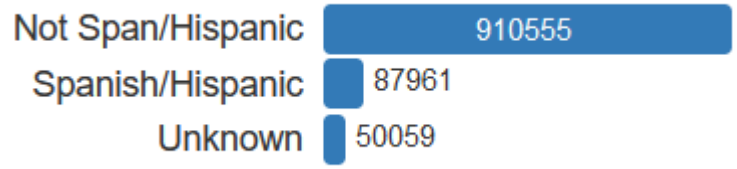
Result :

1 – Genuine Claim

0 -- Fraud Claim

Basic visualizations on attributes





Cultural Group :

Different Cultural Groups like

Black are called as African Americans.

White who are white American

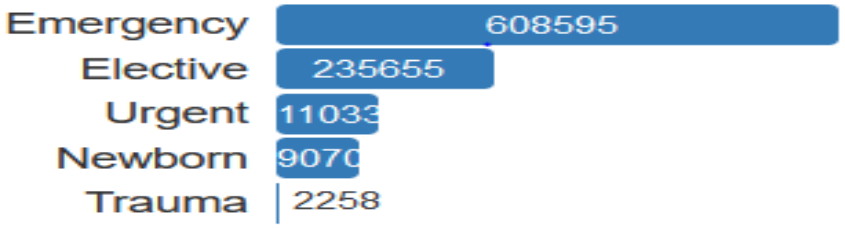
Other Race – Groups other than the above two

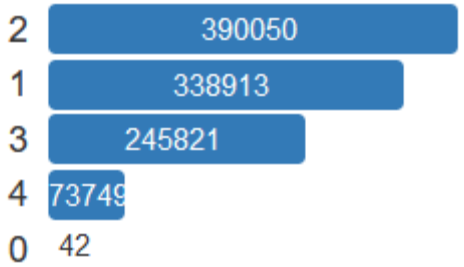
Ethnicity :

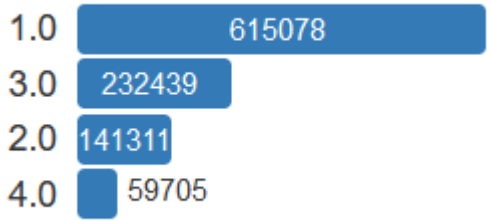
Spanish/Hispanic: Spanish speaking countries especially people from Central and South America

Non-Spanish: Main English speaking countries called Anglo Americans

Basic visualizations on attributes







Admission type:

Different Admission type like Emergency, Elective, Urgent, Newborn, Trauma and not-available

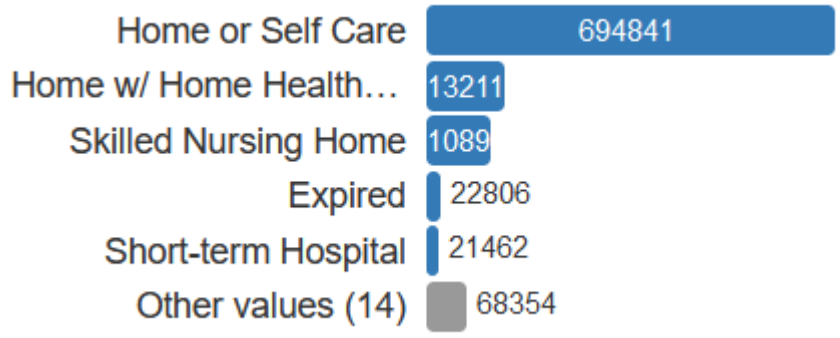
Illness Code :

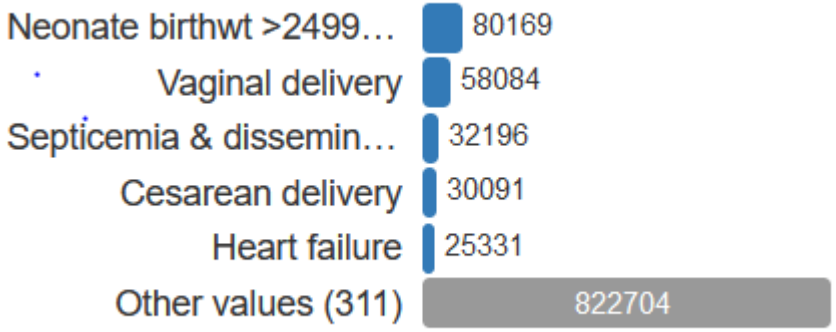
- **0 for stage unspecified**
- **1 for mild**
- **2 for moderate**
- **3 for severe**
- **4 for indeterminate**

Mortality Risk :

- **1 for Minor**
- **2 for Moderate**
- **3 for Major**
- **4 for Severe**

Basic visualizations on attributes





9-Home or Self Care, 10-ccs diagnosis code, 11-ccs procedure code, and 12-APR-DRG (All Patient Refined Diagnosis Related Groups are related to each other according to CCS(Clinical Classification Software) guideline attached.

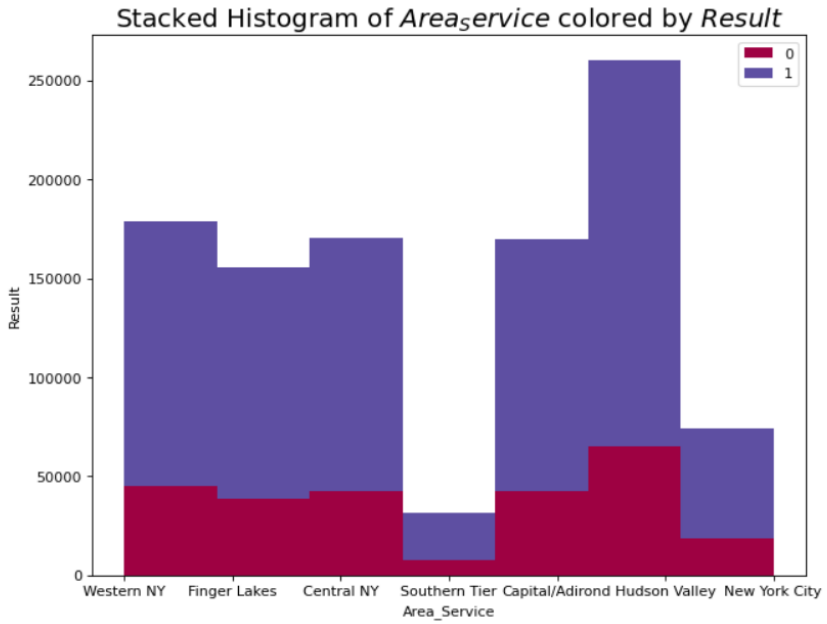
Type of recommended care is covered under Home or Self-care

ccs diagnosis code is the code depending on the type of disease

ccs procedure code is the code for the recommended clinical procedures to be followed.

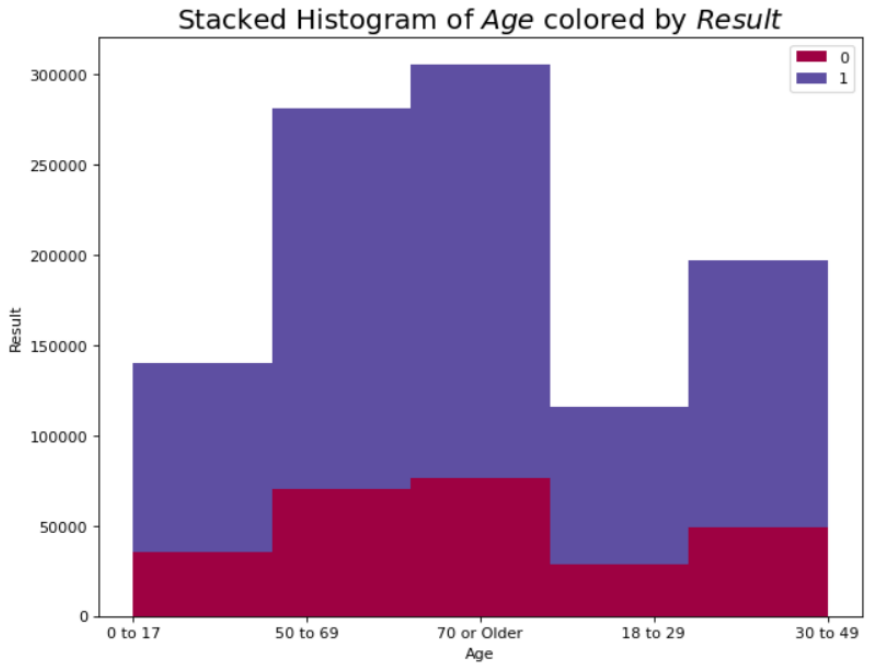
Apr-drg description is the description of the disease

Stacked histogram (Input parameters Vs Output parameter [Result])



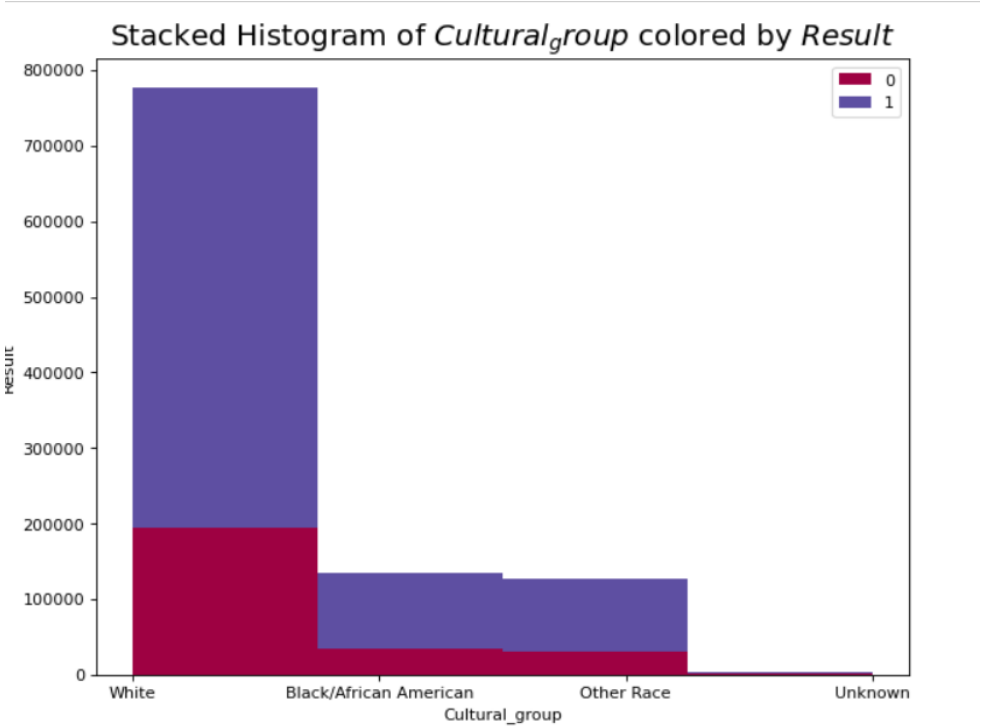
On neglecting Hudson Valley (highest) & Southern Tier (lowest), here adding the fraud claims of rest area services is more than the total claims of New York City.

Stacked histogram (Input parameters Vs Output parameter [Result])



Represents the claim types with resp.to age groups. By looking into the columns individually for each age groups, the claim data is approximately split in such a way that for each respective age groups holds 20-30% of fraud claims and 60-70% of genuine claims of its total claims respectively.

Stacked histogram (Input parameters Vs Output parameter [Result])

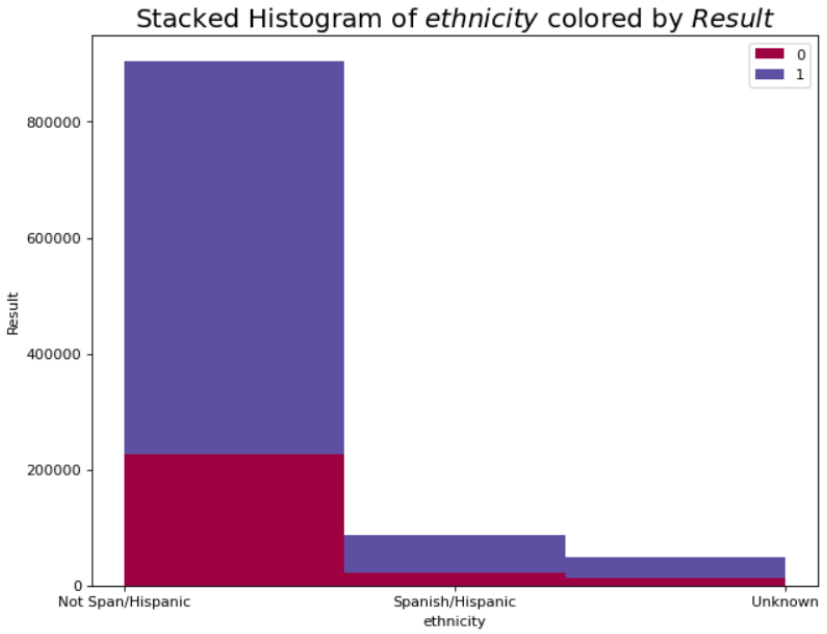


Represents the claim types with resp.to Cultural groups.

Fraud claims of white group exceeds the total claims made by other cultural groups.

Genuine claims of white group exceeds the sum of the total claims made by other cultural groups.

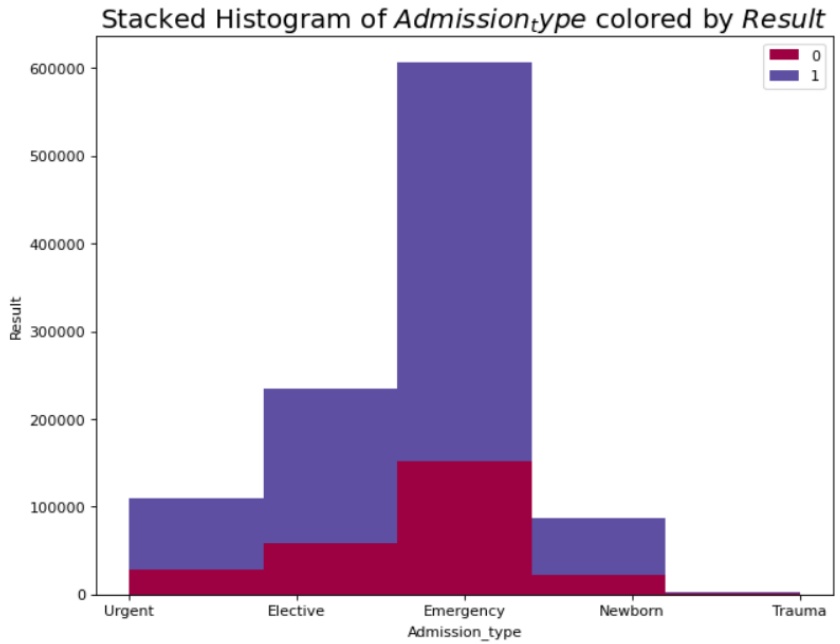
Stacked histogram (Input parameters Vs Output parameter [Result])



Represents the claim types with resp.to ethnicity groups.

Fraud claims of non span group exceeds the sum of the total claims made by other ethnicity groups.

Stacked histogram (Input parameters Vs Output parameter [Result])

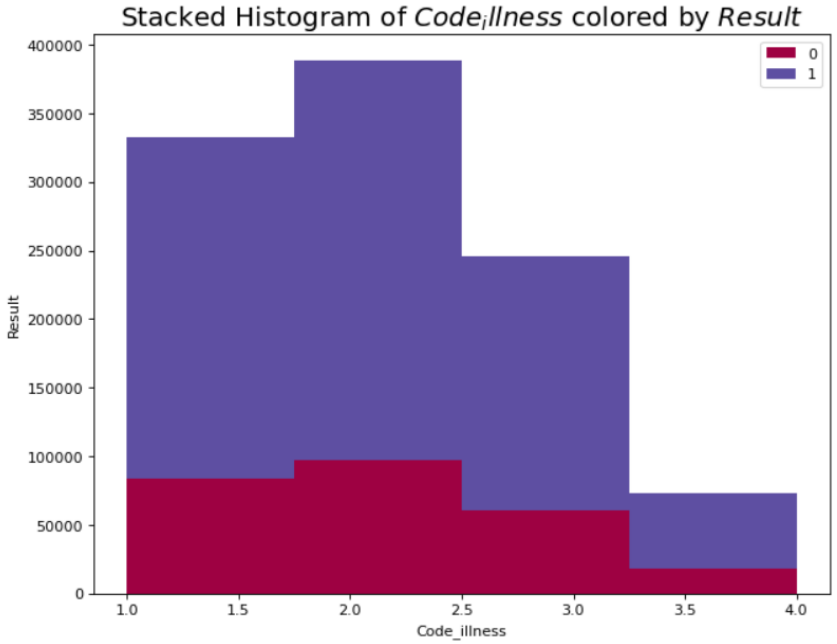


Represents the claim types with resp.to admission types

Fraud claims of Emergency admission exceeds the total claims made by both Urgent and New born admission.

Genuine claims of Emergency admission almost equal to the sum of the total claims made by other admission types.

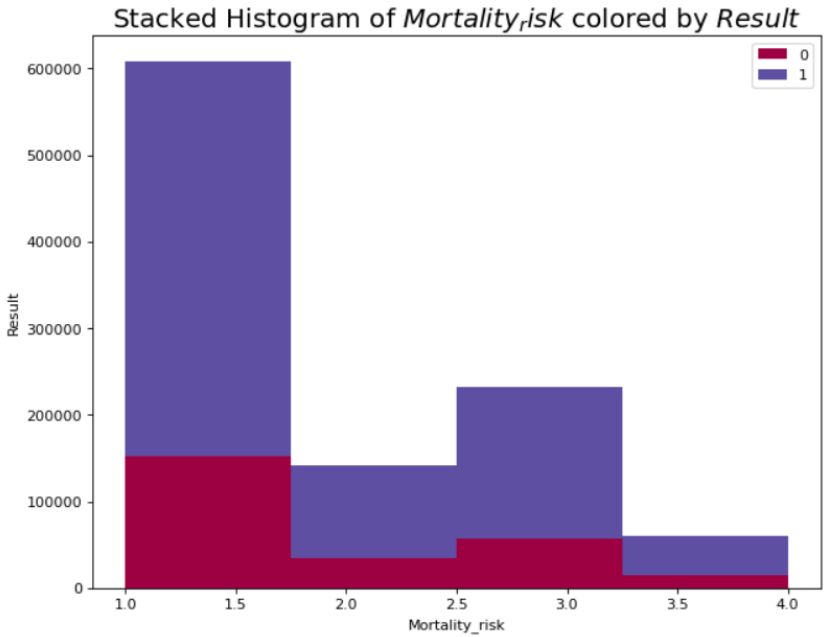
Stacked histogram (Input parameters Vs Output parameter [Result])



Represents the claim types with resp.to Code illness

Fraud claims of 1, 2, 3 almost equal to the total claims made by 4 code illness.

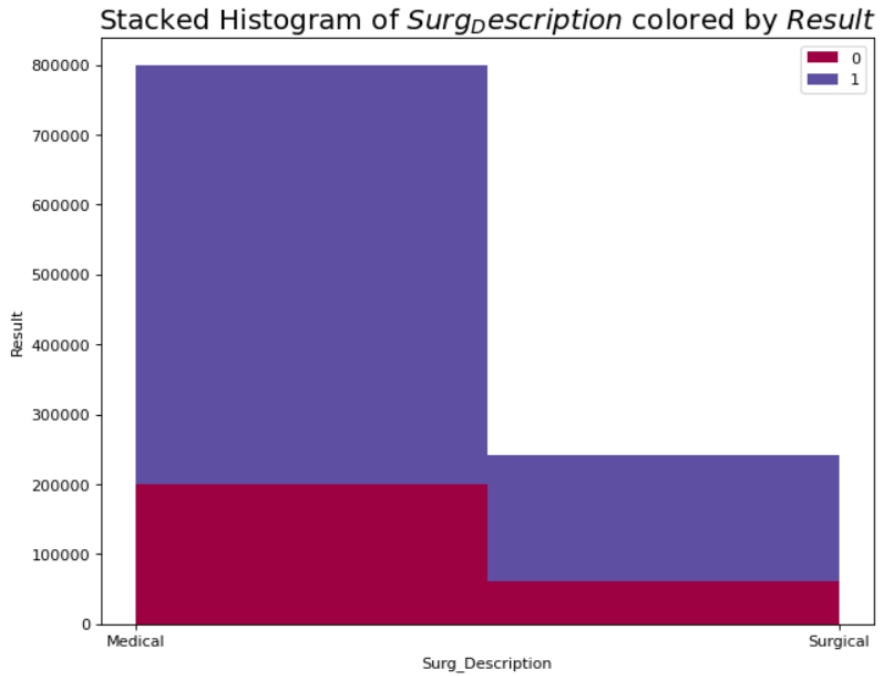
Stacked histogram (Input parameters Vs Output parameter [Result])



Represents the claim types with resp.to Mortality Risk.

Total claims of 1 is almost equal to the sum of the total claims made by other Mortality Risk (2,3,4) groups.

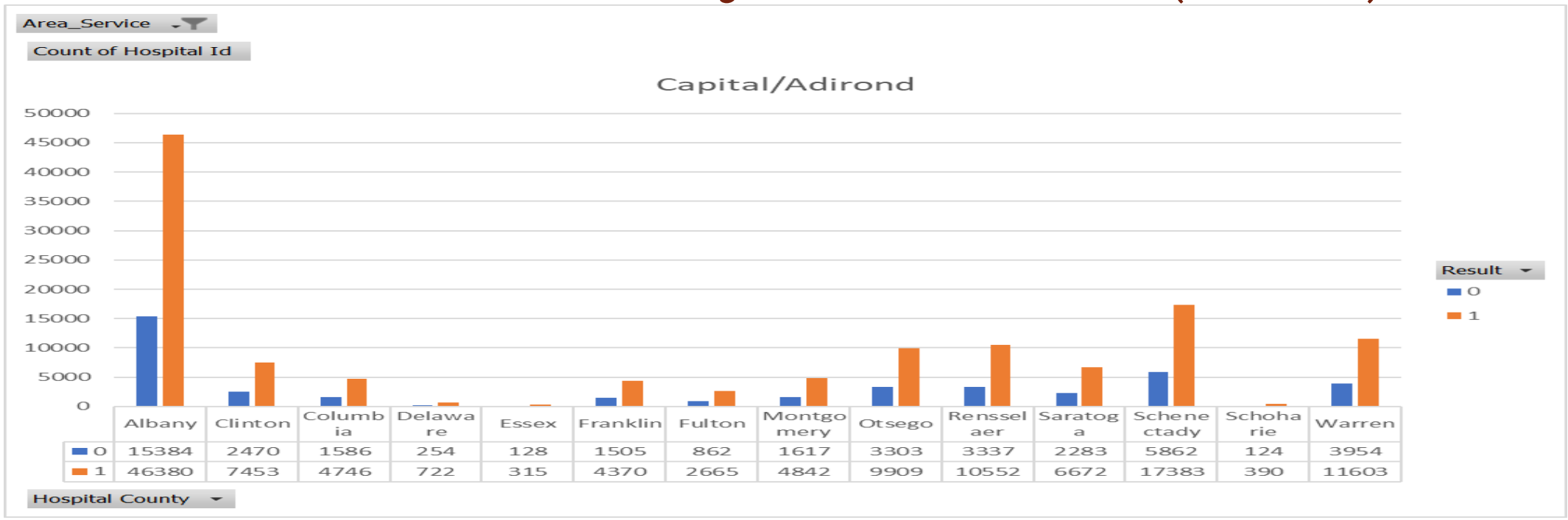
Stacked histogram (Input parameters Vs Output parameter [Result])



Surge_ Description – Type of Treatment – May be Surgical or Medical.

Fraud claims of Medical treatment is almost equal to the total values of surgical treatment.

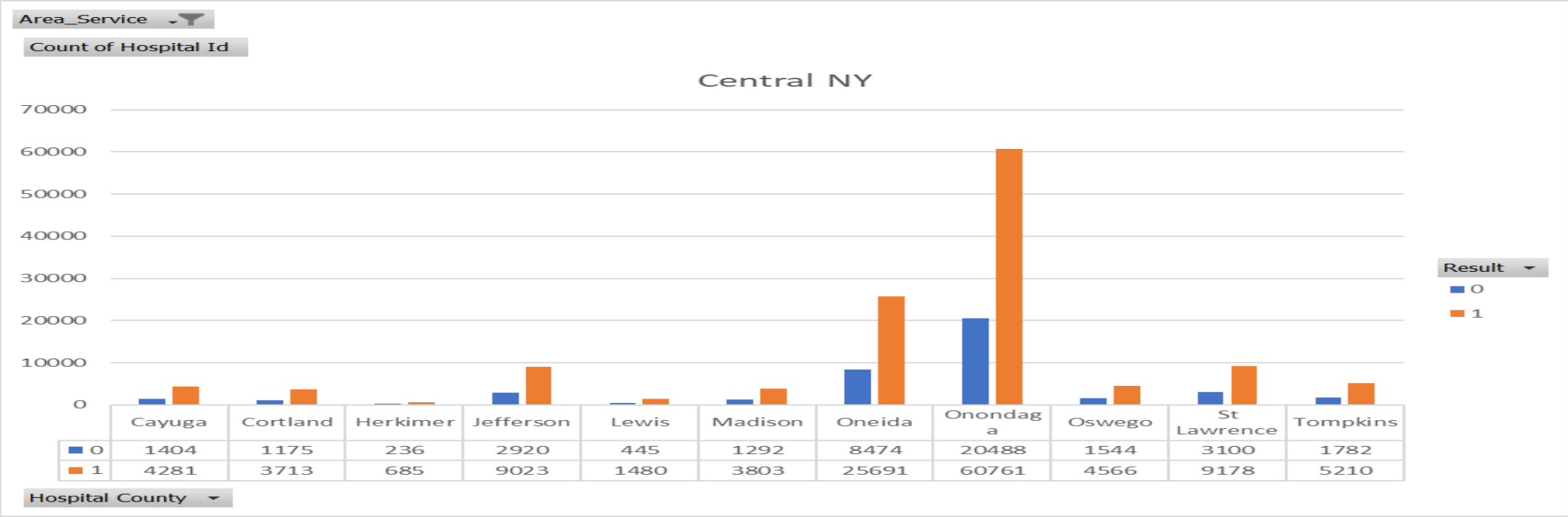
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Highest: Albany
2. Fraud claims of Albany is almost higher than the rest claims when compared with remaining hospital county of Capital/Adirond

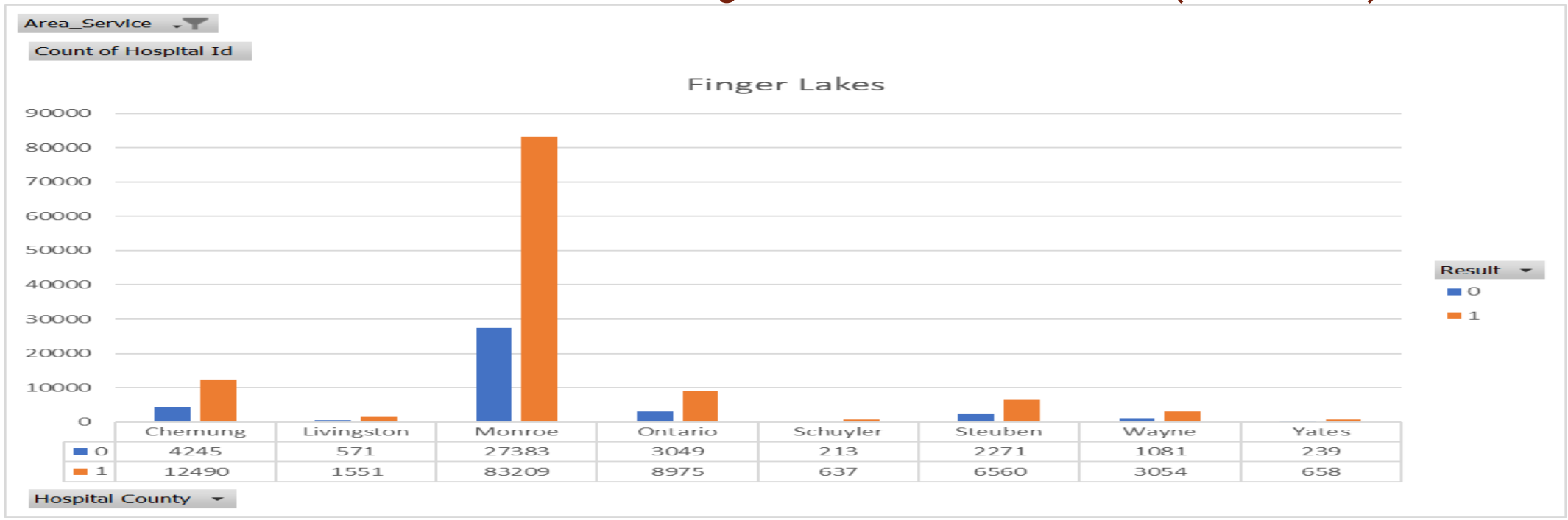
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Highest: Onondaga
2. Fraud claims of Onondaga is almost higher than the rest claims when compared with remaining hospital county.

Visualizations by Pivot chart (Excel)

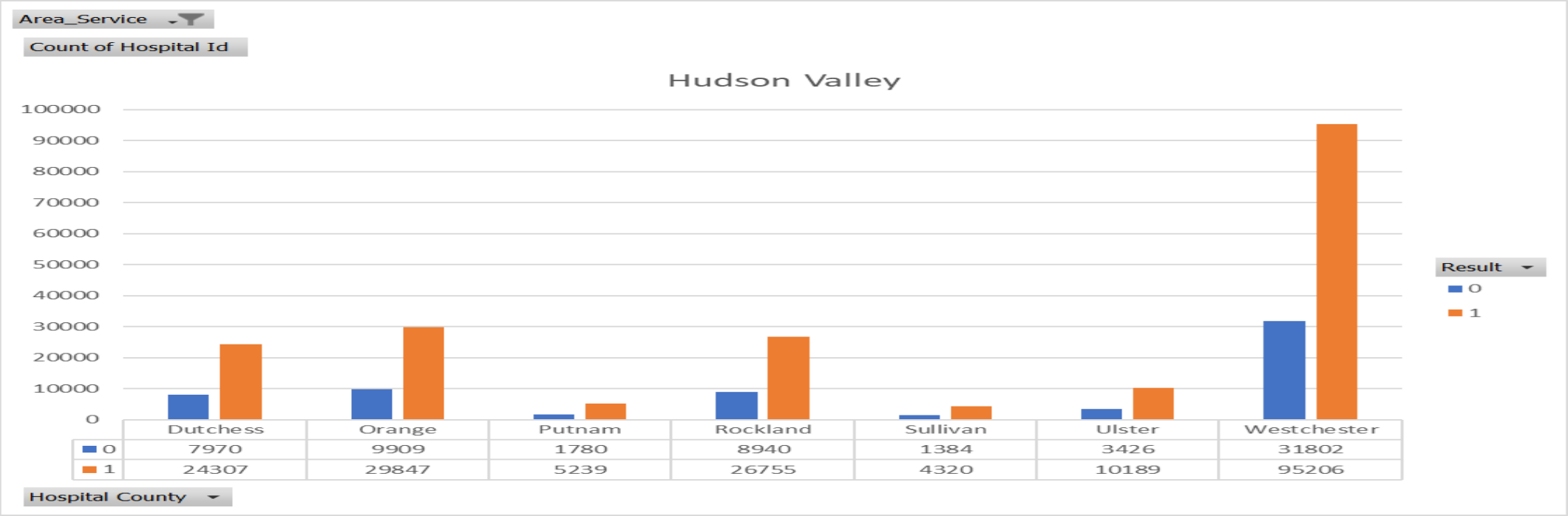


	Chemung	Livingston	Monroe	Ontario	Schuyler	Steuben	Wayne	Yates
0	4245	571	27383	3049	213	2271	1081	239
1	12490	1551	83209	8975	637	6560	3054	658

Hospital county Vs Counts of Hospital ID

1. Highest: Monroe
2. Total Claims of Monroe is the most significant hospital county in finger lakes that it represents the whole data.

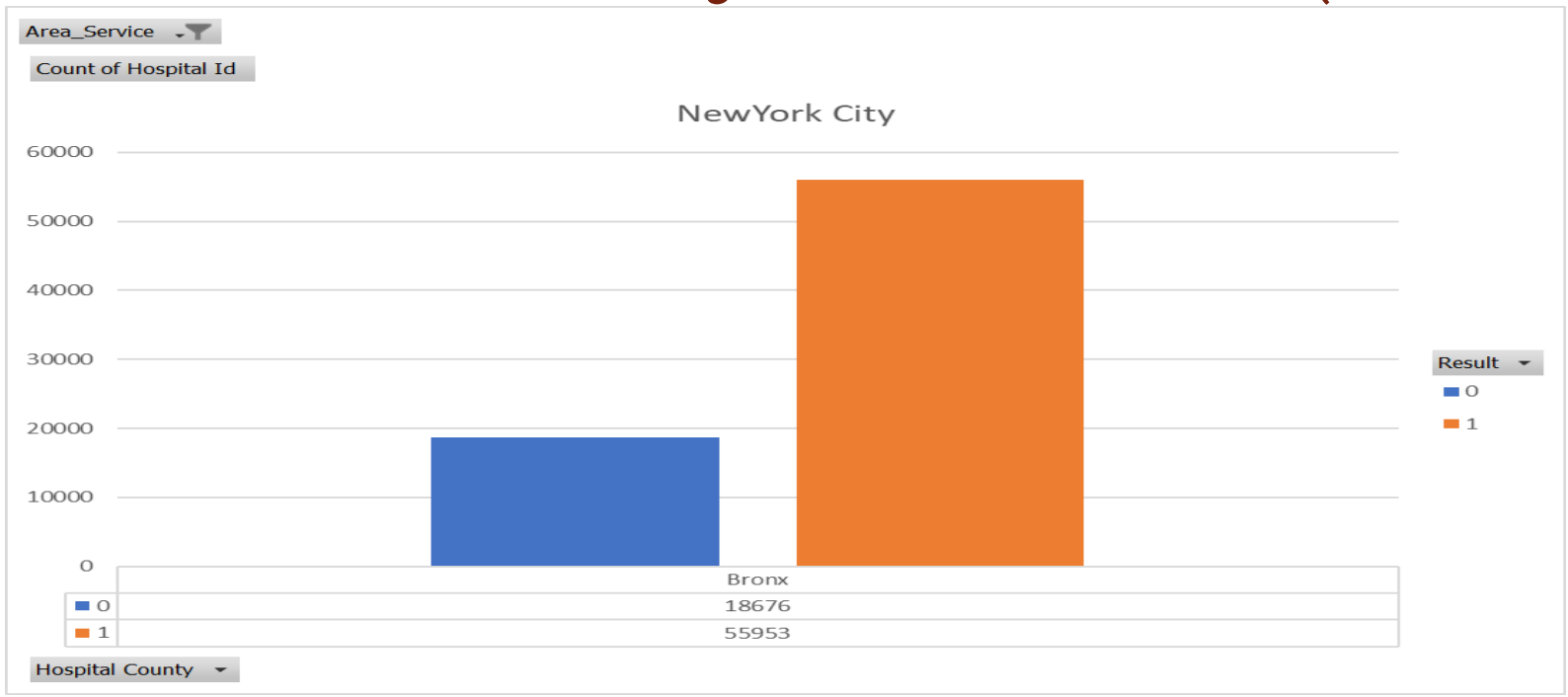
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Highest: Westchester
2. Fraud claims of Westchester is almost higher than the rest claims when compared with remaining hospital county.

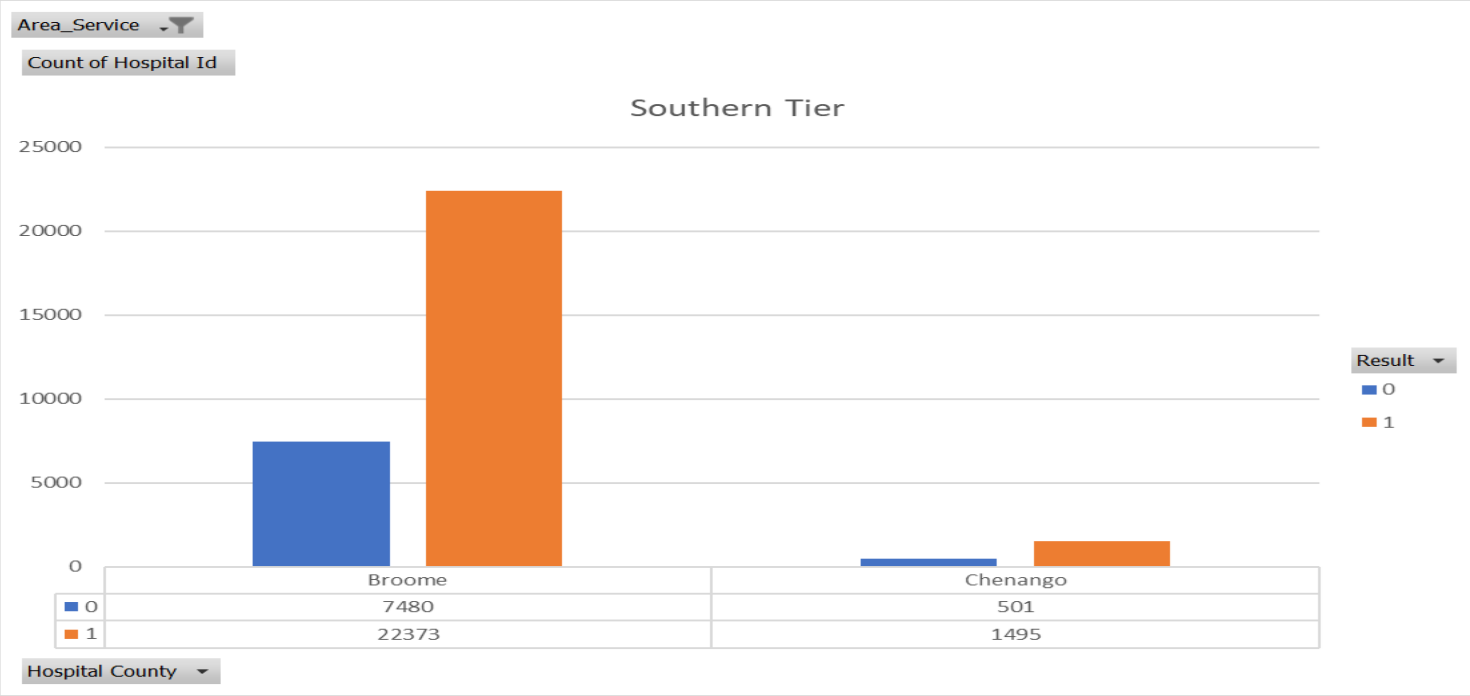
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Only value is Bronx.
2. Total Claims of Bronx is the most significant hospital county in New York City that it represents the whole data.

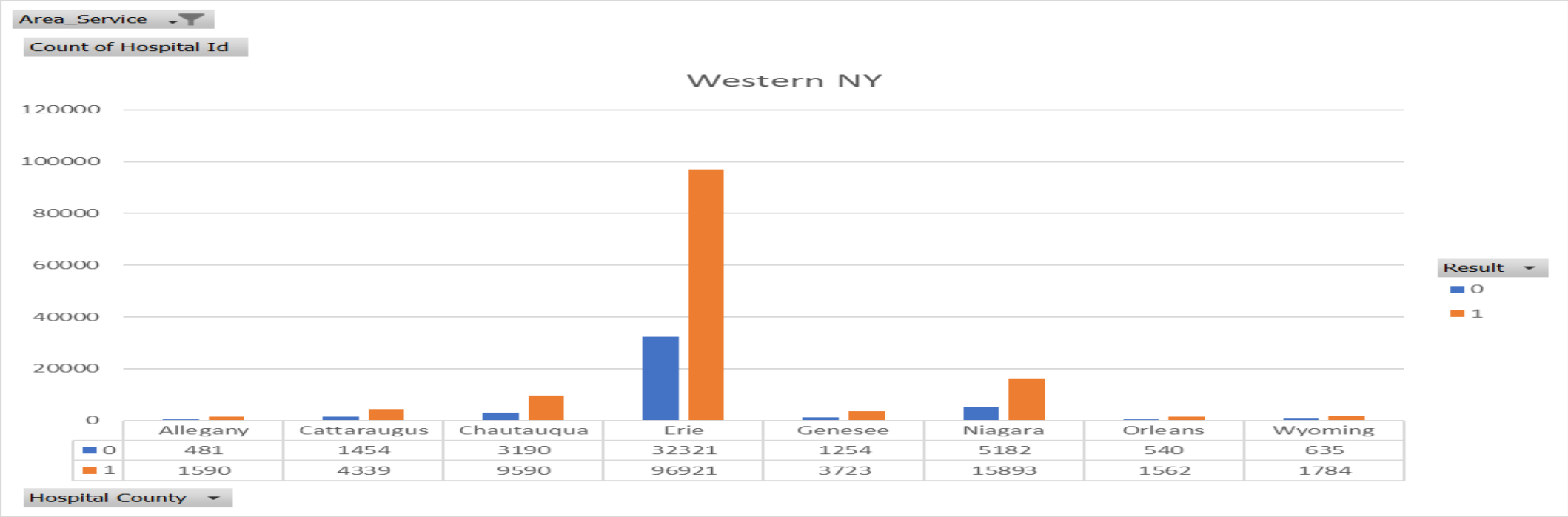
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Only value is Broome & Chenango.
2. Total Claims of Broome is the most significant hospital county in Southern Tier that it represents almost the whole data.

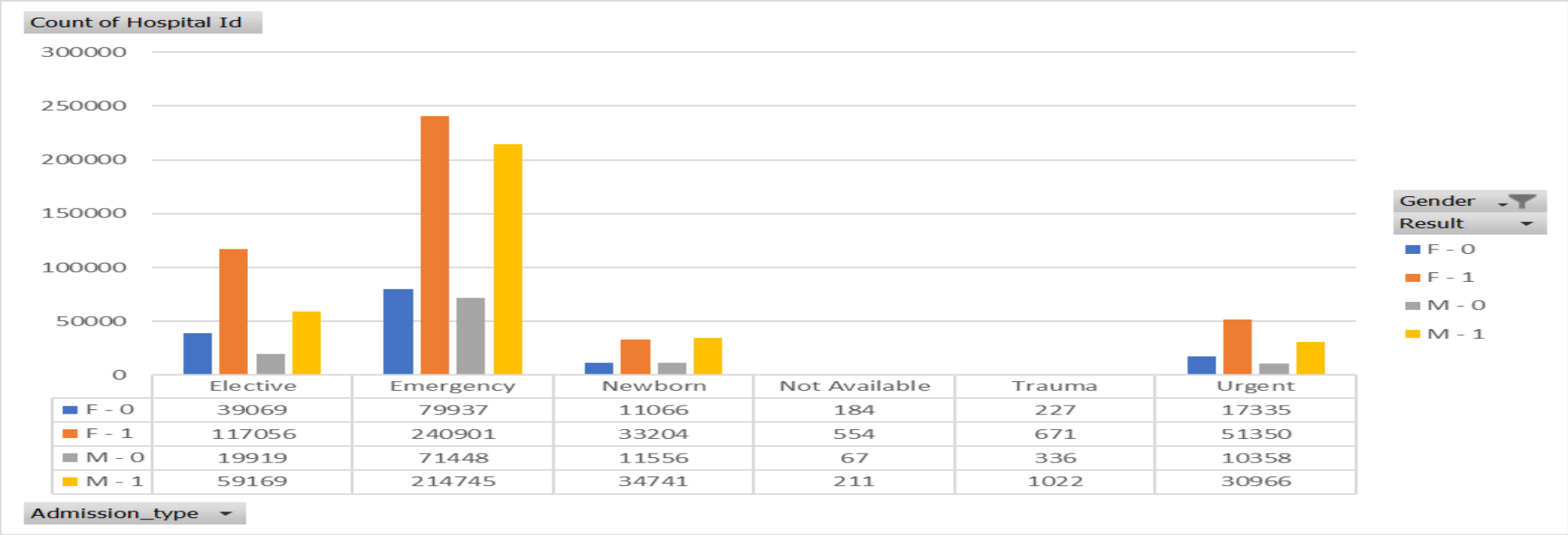
Visualizations by Pivot chart (Excel)



Hospital county Vs Counts of Hospital ID

1. Highest: Erie
2. Genuine claims of Niagara is almost higher than the rest claims when compared with remaining hospital county of Western NY.

Visualizations by Pivot chart (Excel)

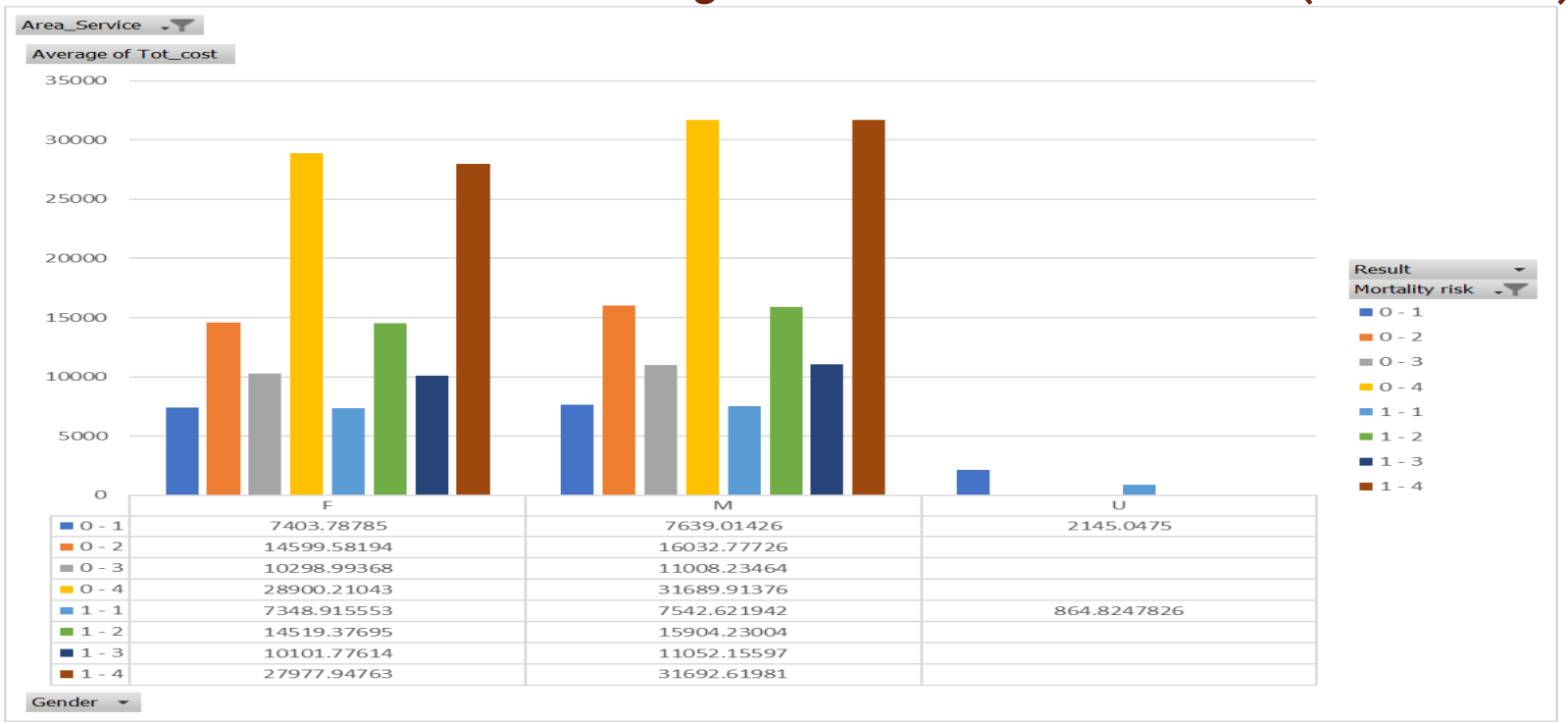


	Elective	Emergency	Newborn	Not Available	Trauma	Urgent
F - 0	39069	79937	11066	184	227	17335
F - 1	117056	240901	33204	554	671	51350
M - 0	19919	71448	11556	67	336	10358
M - 1	59169	214745	34741	211	1022	30966

Admission type Vs Counts of Hospital ID

1. Emergency Admission type holds the majority interaction in the dataset comparitively.
2. Representation shows comparisons of gender with respect to admission types

Visualizations by Pivot chart (Excel)

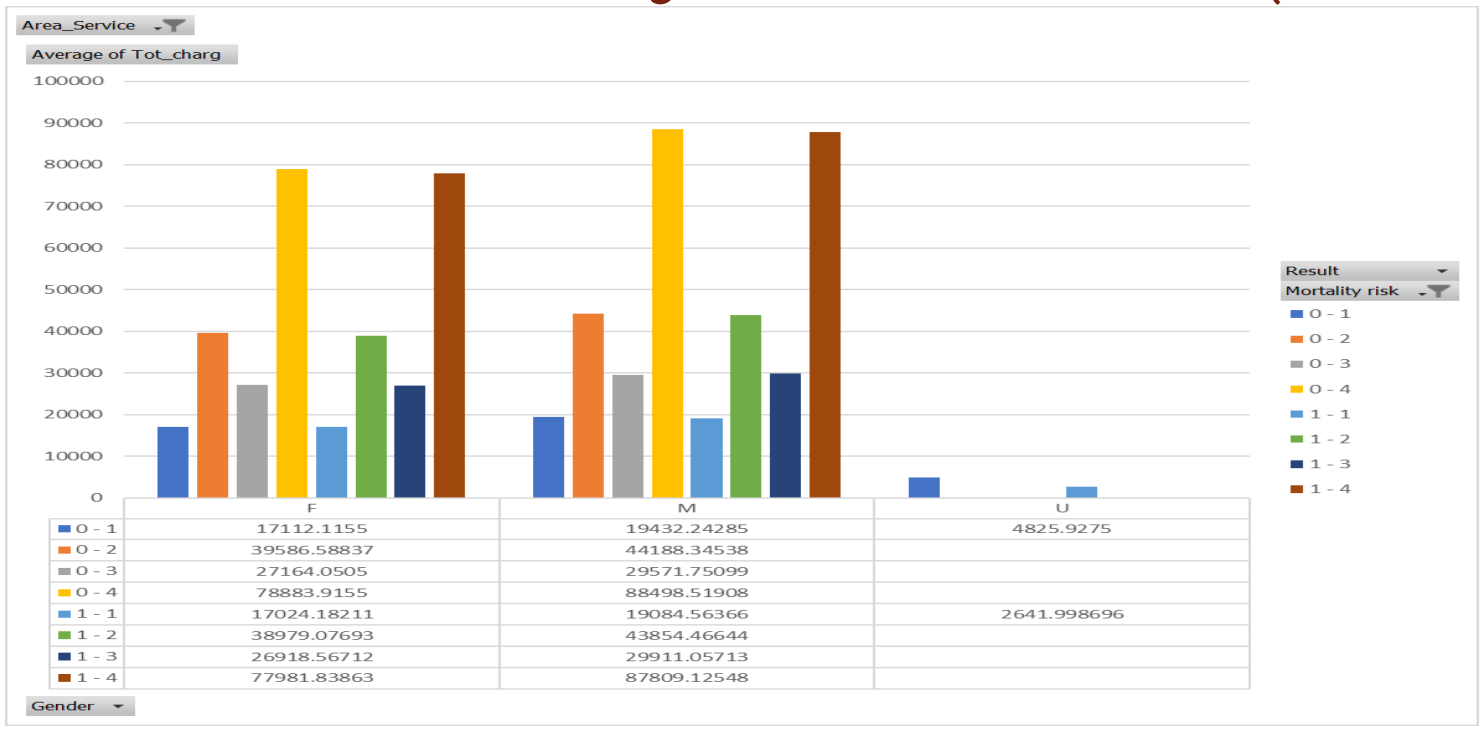


	F	M	U
0 - 1	7403.78785	7639.01426	2145.0475
0 - 2	14599.58194	16032.77726	
0 - 3	10298.99368	11008.23464	
0 - 4	28900.21043	31689.91376	
1 - 1	7348.915553	7542.621942	864.8247826
1 - 2	14519.37695	15904.23004	
1 - 3	10101.77614	11052.15597	
1 - 4	27977.94763	31692.61981	

Gender Vs Average of Total cost

1. Though female made more claims than male, when comparing the average of total cost spend by them, Male tends to spend more (slight above) than female (Both claim types) in the dataset.
2. Representation shows comparisons of gender with respect to Average total cost

Visualizations by Pivot chart (Excel)

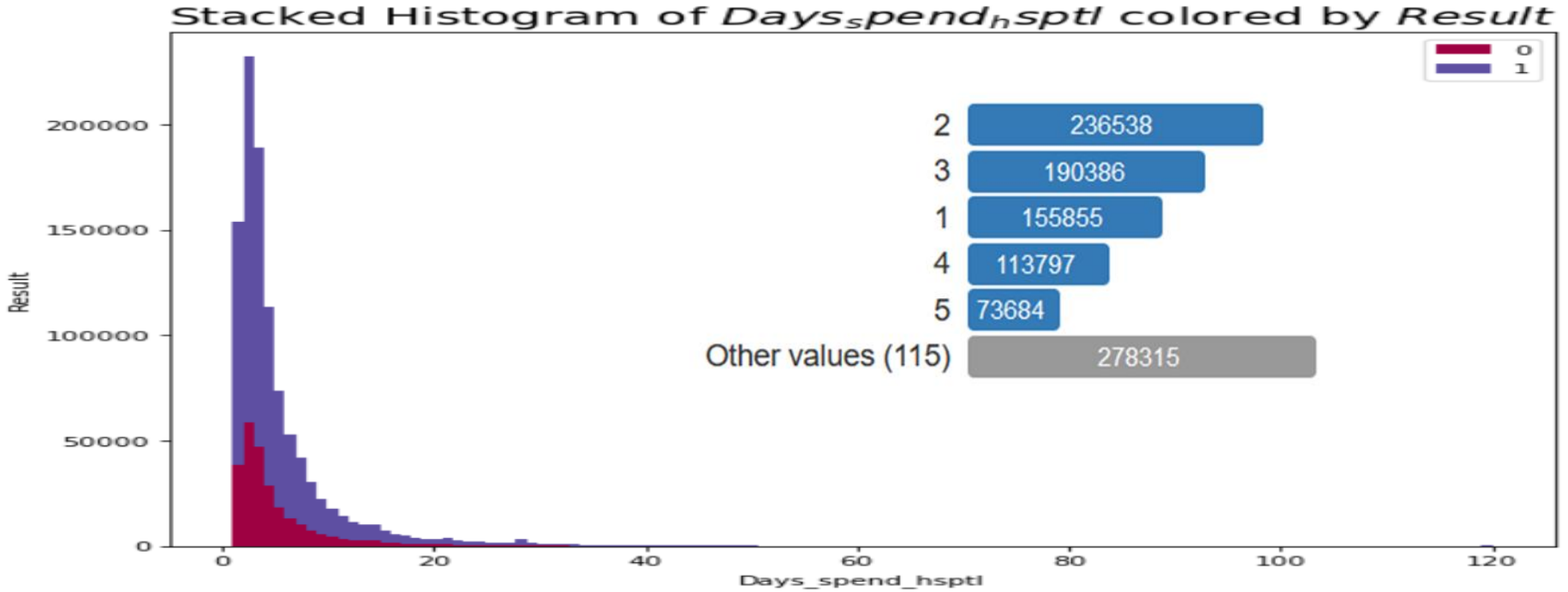


	F	M	U
0 - 1	17112.1155	19432.24285	4825.9275
0 - 2	39586.58837	44188.34538	
0 - 3	27164.0505	29571.75099	
0 - 4	78883.9155	88498.51908	
1 - 1	17024.18211	19084.56366	2641.998696
1 - 2	38979.07693	43854.46644	
1 - 3	26918.56712	29911.05713	
1 - 4	77981.83863	87809.12548	

Gender Vs Average of Total charge

1. Though female made more claims than male, when comparing the average of total charge spend by them, hence in most cases Male tends to spend more (slight above) than female in the dataset.
2. Representation shows comparisons of gender with respect to Average total charge.

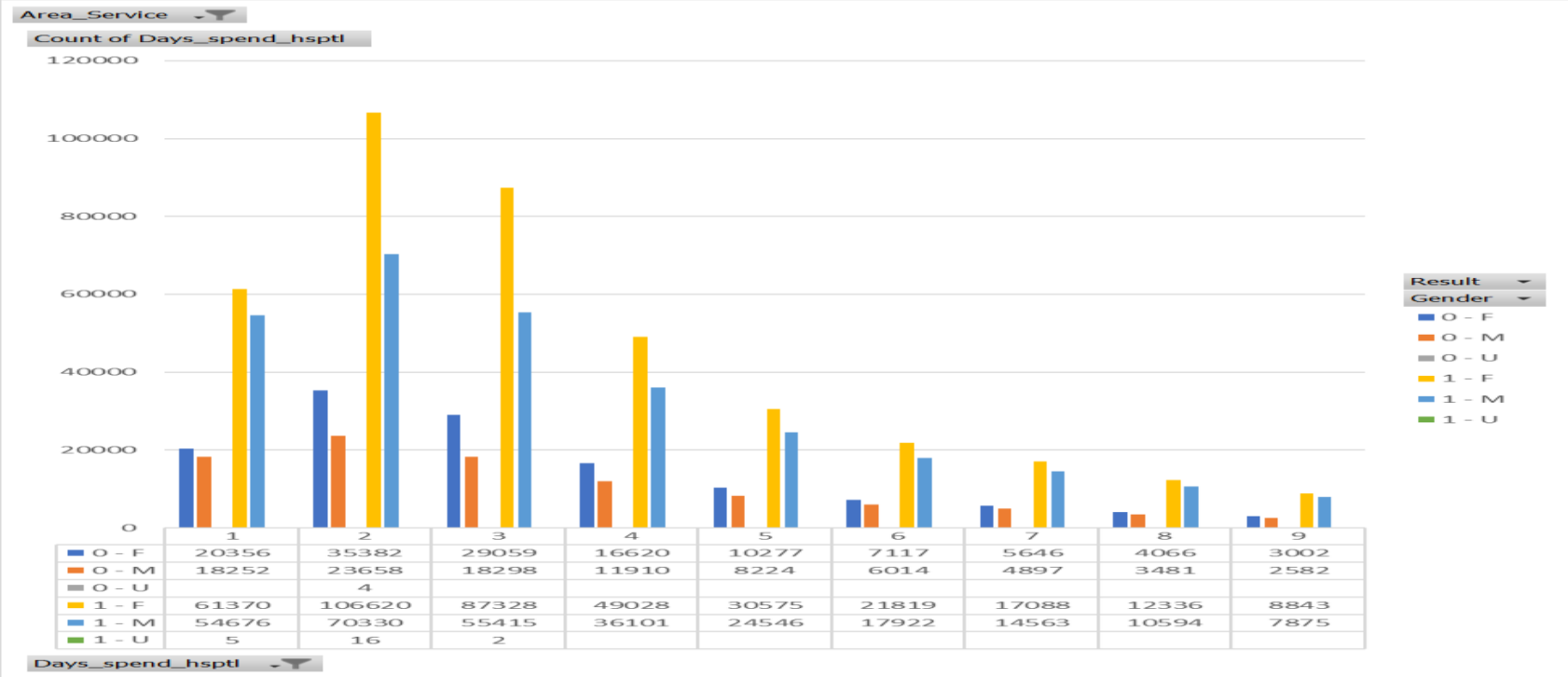
Visualizations by Pivot chart (Excel)



Days spent in hospital

1. Representation shows counts of days spent in hospital.
2. Data is more skewed from 1 to 20 days among 120 unique days and here there are more cases spends 2 days mandatory in the hospital. (Top 5 data Vs Other values)
3. Values of 2 days spent in hospital is almost close to other 115 values.

Visualizations by Pivot chart (Excel)

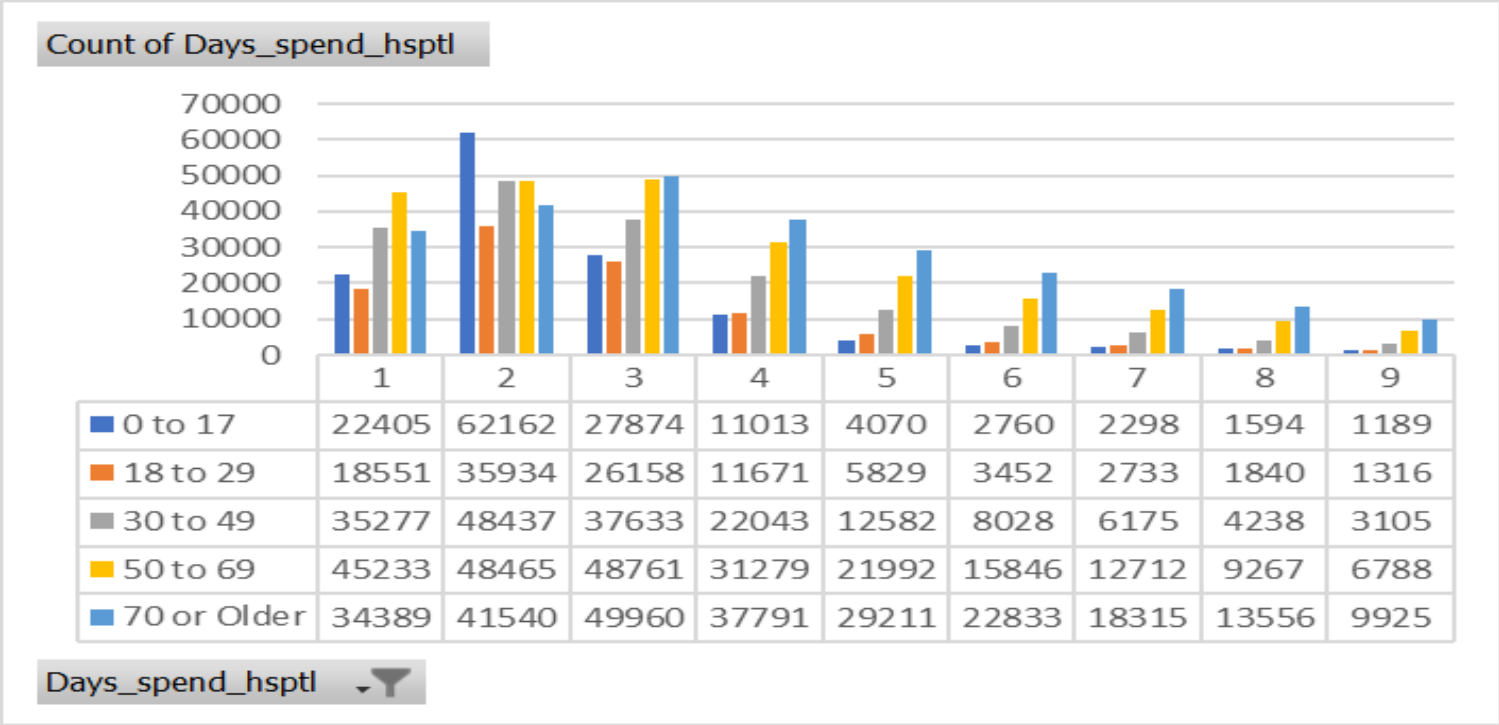


	1	2	3	4	5	6	7	8	9
0 - F	20356	35382	29059	16620	10277	7117	5646	4066	3002
0 - M	18252	23658	18298	11910	8224	6014	4897	3481	2582
0 - U		4							
1 - F	61370	106620	87328	49028	30575	21819	17088	12336	8843
1 - M	54676	70330	55415	36101	24546	17922	14563	10594	7875
1 - U	5	16	2						

Days spent in hospital

1. Representation shows counts of days spent in hospital (Top 9) Vs. Gender
2. As the days spent in hospital increases we see 20-25% difference between claim counts by gender in the beginning and as further the days spend increases more than 10 days we see that difference is reducing among the genders.

Visualizations by Pivot chart (Excel)



	1	2	3	4	5	6	7	8	9
0 to 17	22405	62162	27874	11013	4070	2760	2298	1594	1189
18 to 29	18551	35934	26158	11671	5829	3452	2733	1840	1316
30 to 49	35277	48437	37633	22043	12582	8028	6175	4238	3105
50 to 69	45233	48465	48761	31279	21992	15846	12712	9267	6788
70 or Older	34389	41540	49960	37791	29211	22833	18315	13556	9925

Age Vs Days spent in hospital

1. Representation shows counts of days spent in hospital (Top 9) Vs. Age Groups
2. 50-69 Age group shows high as they might be coming in for a single day for General check up
3. 0-17 Age group; Among all the rest data, this group alone differs from regular pattern only in Days spent in hospital (2days) and stands highest when compared to the rest all data.