Hospital Mortality Prediction
(Milestone 3)

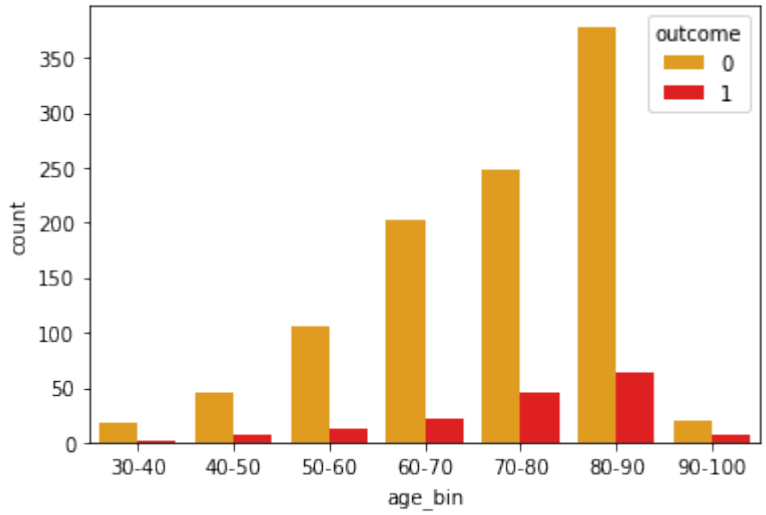
Data Pre-processing:

#1. Null values detection and treatment
- 'outcome' and 'Pulse rate cat' were having 1 and 16 null respectively
- Imputed with mode(as both were categorical features).

#2. Outliers detection and treatment
- Detected outliers in 'age' column with IQR method
- Since this is a medical dataset and few records were there(1177).....each record Is valuable.
- So treated the outliers by capping method.

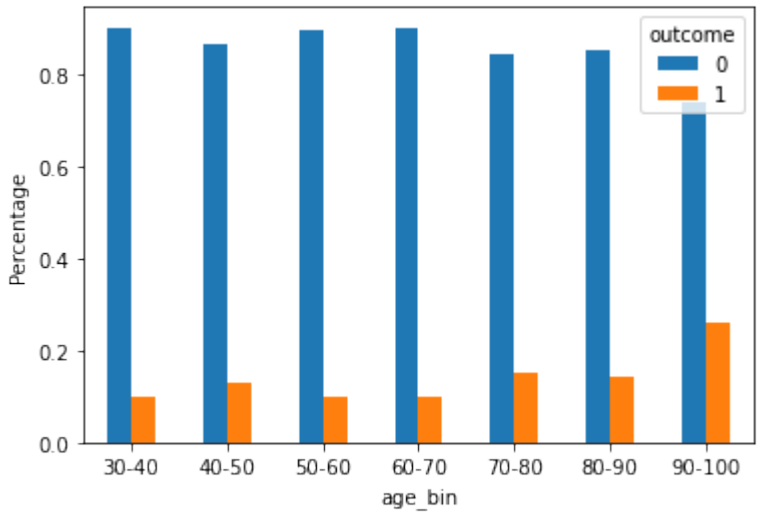
Feature Engineering:

#1. 'age_bin' feature
- Transformed 'age' column to 'age_bin' by creating bins of 10y each from 30y to 100y.
- Intention was to see weather age has any affect on mortality.
- 'age_bin' p-value: 0.18964
- We retain the H0, as there is no relation between 'age_bin' and 'outcome'
- Count Plot



Here, 0: Both are out of range, 1: Either of them is out of range, 2: Both are in range

- Percentage Graph



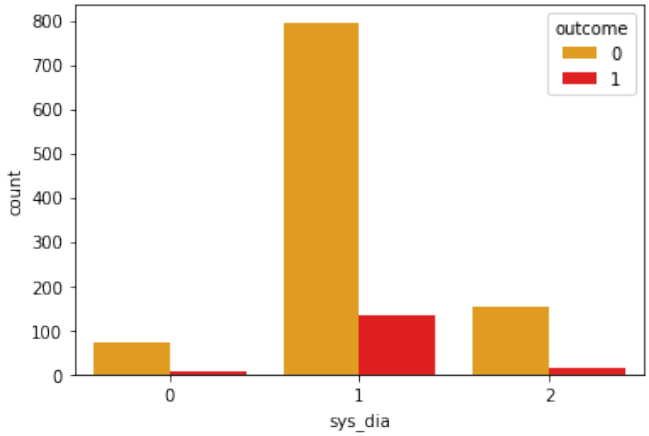
Here, 0: Both are out of range, 1: Either of them is out of range, 2: Both are in range

Cross Tab

outcome	0	1	Percent(%)
age_bin			
30-40	18	2	10.000000
40-50	46	7	13.207547
50-60	106	12	10.169492
60-70	202	22	9.821429
70-80	248	45	15.358362
80-90	378	64	14.479638
90-100	20	7	25.925926

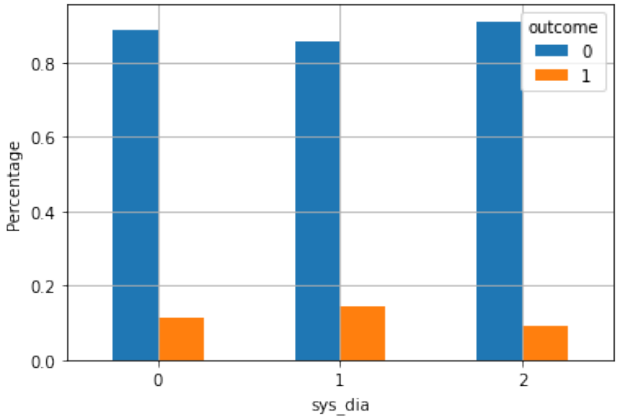
#2. ‘sys_dia’ feature

- Clubbed both ‘Sys_cat’ and ‘Diastolic’ as both features are components of blood pressure
- Intention was to see weather this new ‘sys_dia’ feature has any affect on mortality.
- ‘sys_dia’ p-value: 0.12965
- We retain the H0, as there is no relation between ‘sys_dia’ and ‘outcome’
- Count Plot



Here, 0: Both are out of range, 1: Either of them is out of range, 2: Both are in range

- Percentage Graph



Here, 0: Both are out of range, 1: Either of them is out of range, 2: Both are in range

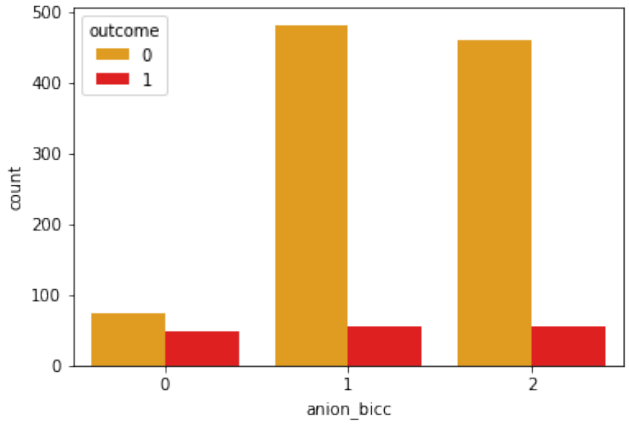
- Cross Tab

outcome	0	1	Percent(%)
sys_dia			
0	71	9	11.250000
1	795	135	14.516129
2	152	15	8.982036

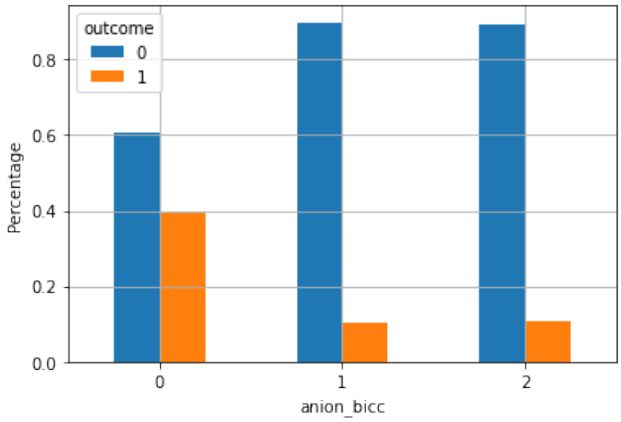
#3. ‘anion_bicc’ feature

- Clubbed both ‘anion_cat’ and ‘Biccarbon_cat’ as both features showcase acid-base behaviour.
- Intention was to see weather this new ‘anion_bicc’ feature has any affect on mortality.
- ‘anion_bicc’ p-value: 0.00
- We reject the H0, as there is a relation between ‘anion_bicc’ and ‘outcome’

- Count Plot



- Percentage Graph



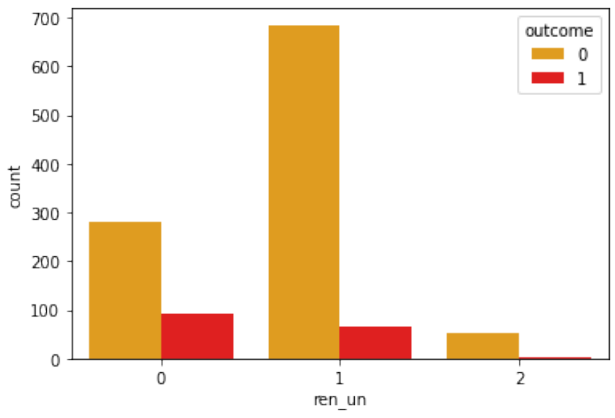
- Cross Tab

outcome	0	1	Percent(%)
anion_bicc			
0	75	49	39.516129
1	482	55	10.242086
2	461	55	10.658915

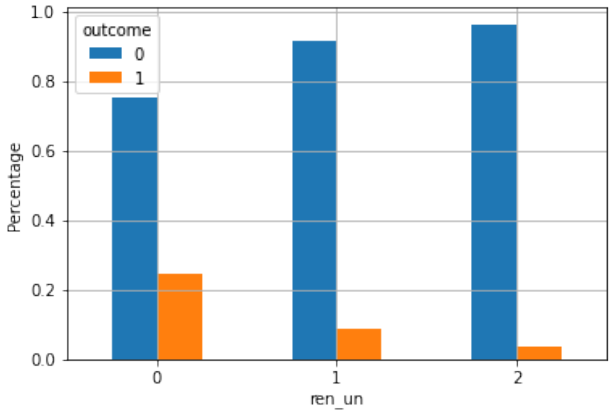
#4. 'ren_un' feature

- Clubbed both 'Renal_failure' and 'UN_cat' as both features showcase problem with kidney function.
- Intention was to see weather this new 'ren_un' feature has any affect on mortality.
- 'ren_un' p-value: 0.00
- We reject the H0, as there is a relation between 'ren_un' and 'outcome'

- Count Plot



- Percentage Graph

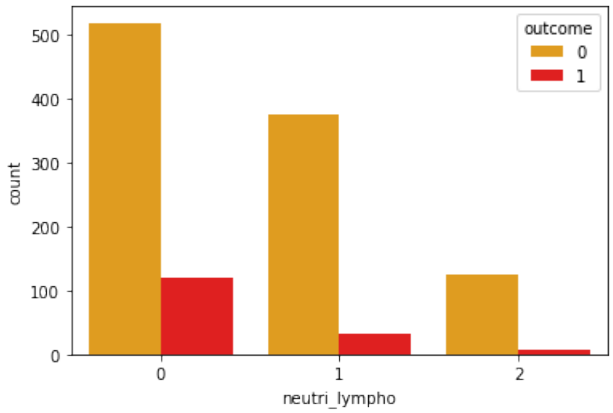


- Cross Tab

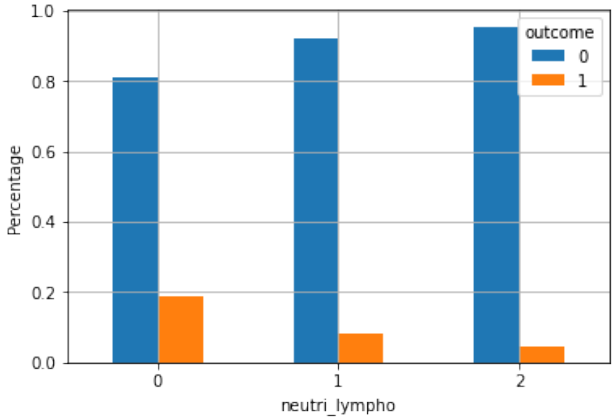
outcome	0	1	Percent(%)
ren_un			
0	282	92	24.598930
1	685	65	8.666667
2	51	2	3.773585

#5. 'neutri_lympho' feature

- Clubbed both 'neutriphil_cat' and 'Lympho_cat' as both features showcase a type of WBC count.
- Intention was to see weather this new 'neutri_lympho' feature has any affect on mortality.
- 'neutri_lympho' p-value: 0.00
- We reject the H0, as there is a relation between 'neutri_lympho' and 'outcome'
- Count Plot



- Percentage Graph



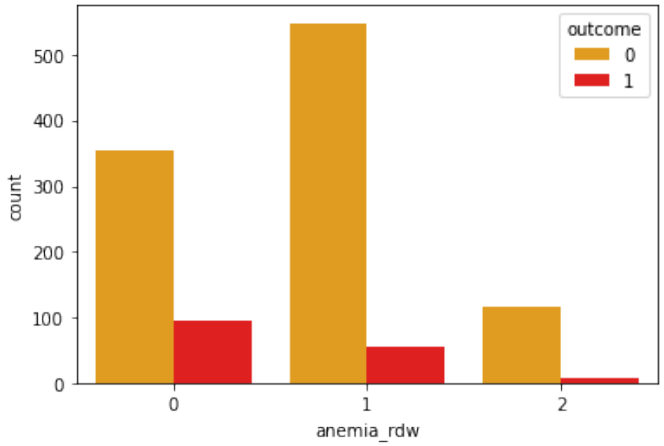
- Cross Tab

outcome	0	1	Percent(%)
neutri_lympho			
0	518	120	18.808777
1	376	33	8.068460
2	124	6	4.615385

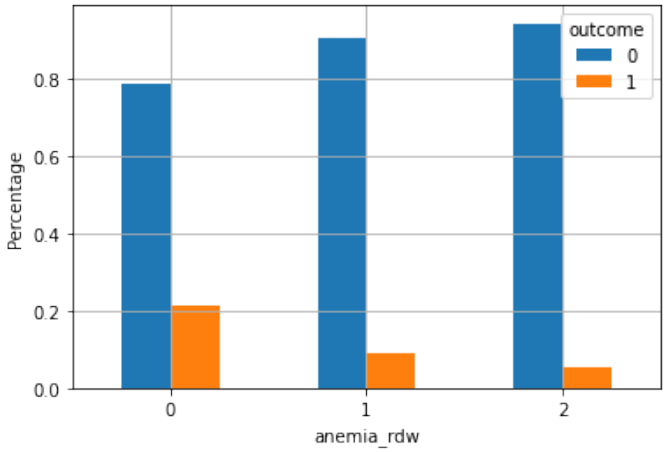
#5. ‘anemia_rdw’ feature

- Clubbed both ‘deficiencyanemias’ and ‘rdw_cat’ as both features showcase a type of WBC count.
- Intention was to see weather this new ‘anemia_rdw’ feature has any affect on mortality.
- ‘anemia_rdw’ p-value: 0.00
- We reject the H0, as there is a relation between ‘anemia_rdw’ and ‘outcome’

- Count Plot



- Percentage Graph



- Cross Tab

outcome	0	1	Percent(%)
anemia_rdw			
0	353	96	21.380846
1	548	56	9.271523
2	117	7	5.645161

Removing unwanted features:

- We will remove all the features which are having p-Value greater than 0.05 (features where H0 is getting retained)
- Also we will remove all those features which were used for new feature creation (those new features which were selected for model training).

- Final selected features:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1177 entries, 0 to 1176
Data columns (total 27 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   outcome             1177 non-null   int64
 1   hypertensive        1177 non-null   int64
 2   atrialfibrillation  1177 non-null   int64
 3   depression          1177 non-null   int64
 4   Pulse rate cat      1177 non-null   int64
 5   Sys_cat             1177 non-null   int64
 6   Diastolic           1177 non-null   int64
 7   respiratory cat     1177 non-null   int64
 8   temp_cat            1177 non-null   int64
 9   urine_cat           1177 non-null   int64
 10  mcv_cta             1177 non-null   int64
 11  leukocytes_cat      1177 non-null   int64
 12  platelets_cat       1177 non-null   int64
 13  Basophil_cat        1177 non-null   int64
 14  PT_cat(sec)         1177 non-null   int64
 15  Creatinine_cat      1177 non-null   int64
 16  potas_cat           1177 non-null   int64
 17  sodium_cat          1177 non-null   int64
 18  cal_cat             1177 non-null   int64
 19  chloride_cat        1177 non-null   int64
 20  ph_cat              1177 non-null   int64
 21  metcat              1177 non-null   int64
 22  pco2_cat            1177 non-null   int64
 23  anion_bicc          1177 non-null   int64
 24  ren_un              1177 non-null   int64
 25  neutri_lympho       1177 non-null   int64
 26  anemia_rdw          1177 non-null   int64
dtypes: int64(27)
memory usage: 248.4 KB
```