Assignment 1

October 13, 2019

1 MScBMI 33200 – Machine Learning for Biomedical Informatics

2 Assignment I

Name: Troy Zhongyi Zhang
Netid: zhongyiz@uchicago.edu

Directions

1. Follow instructions below for each question

2. You can use either R or Python for completing the assignment

3. Upload your answer sheet along with your code (as HTML or PDF) in a separate file. R users can Knit an R markdown into HTML/PDF. Python users can use IPython or Jupyer notebooks and convert them into HTML/PDF (see instructions)

3 Section 1: EMR Bots 30-day Readmission study

3.0.1 Read Data (Preparation)

[1]:
```
import pandas as pd
import numpy as np
from tableone import TableOne
info = pd.read_csv("encounter_info.csv")
labs = pd.read_csv("encounter_labs.csv")
df = pd.read_csv("readmission_outcome.csv")

info.shape
```

[1]:
```
(36143, 7)
```

[2]:
```
labs.shape
```

[2]:
```
(4902804, 18)
```

[3]:
```
df.shape
```

[3]:
```
(36143, 2)
```

4 Section 1, (Q1):

[4]:
```
df1 = pd.merge(info, df, on="Encounter_ID")
```

[5]:
```
df1.head()
```

[5]:
```
   Patient_ID Encounter_ID       AdmissionStartDate         AdmissionEndDate  \
0      100000     100000_2  1996-01-20 14:32:48.440  1996-02-06 17:11:05.247
1      100000     100000_1  1981-01-06 01:54:13.577  1981-01-08 23:35:05.233
2      100000     100000_5  2007-04-12 03:30:56.917  2007-04-27 18:30:58.473
3      100000     100000_3  2002-04-06 23:17:11.963  2002-04-11 00:49:49.810
4      100000     100000_7  2012-07-18 12:01:05.853  2012-07-21 04:17:51.173

  PatientGender PatientRace  PatientEncounterAge  outcome
0        Female       White            36.456455        0
1        Female       White            21.408437        0
2        Female       White            47.688073        0
3        Female       White            42.671151        0
4        Female       White            52.960276        0
```

[6]:
```
columns = ['PatientEncounterAge','PatientGender','PatientRace']
categorical = ['PatientGender','PatientRace']
groupby = 'outcome'
mytable = TableOne(df1, columns, categorical, groupby, pval=True)
print(mytable)
```

```
                                    Grouped by outcome
                                                isnull             0
1   pval              ptest
variable          level
n                                                           36015
128
PatientEncounterAge                                    0   41.7 (18.1)  44.3
(18.0)  0.108  Two Sample T-test
PatientGender     Female                               0  18812 (52.2)    64
(50.0)  0.677      Chi-squared
                  Male                                    17203 (47.8)    64
(50.0)
PatientRace       African American                     0   5382 (14.9)    21
(16.4)  0.580      Chi-squared
                  Asian                                    8251 (22.9)    33
(25.8)
                  Unknown                                  4682 (13.0)    19
(14.8)
                  White                                   17700 (49.1)    55
(43.0)
[1] Warning, Hartigan's Dip Test reports possible multimodal distributions for:
PatientEncounterAge.
[2] Warning, test for normality reports non-normal distributions for:
```

PatientEncounterAge.

[7]:
```
from scipy.stats import ttest_ind
df30 = df1[df1['outcome']==1]
dfnot30 = df1[df1['outcome']==0]
ttest_ind(df30['PatientEncounterAge'], dfnot30['PatientEncounterAge'])
```

[7]:
```
Ttest_indResult(statistic=1.6180629064428116, pvalue=0.10565776793068646)
```

5 Section 1, (Q2):

5.0.1 Find out the latest lab dates for each encounter

[8]:
```
lab=labs.copy()
```

[9]:
```
lab['Lab_DTTM'] = pd.to_datetime(lab['Lab_DTTM'])
```

[10]:
```
grouped_lab = lab.sort_values("Lab_DTTM", ascending=False
                              ).groupby("Encounter_ID")["Lab_DTTM"
                                                        ].apply(list).apply(pd.
 ↪Series
                                                                            ).
 ↪reset_index()
grouped_lab = grouped_lab[["Encounter_ID", 0]]
grouped_lab.columns = ["Encounter_ID", "Lab_DTTM"]
```

[11]:
```
grouped_lab.shape
```

[11]:
```
(36143, 2)
```

[12]:
```
newlab0 = pd.merge(grouped_lab, lab, on=['Encounter_ID','Lab_DTTM'])
```

[13]:
```
newlab0.shape
```

[13]:
```
(36173, 18)
```

5.0.2 Only keep the last lab data row for each encounter

[14]:
```
newlab0 = newlab0.drop_duplicates('Encounter_ID', keep='last').values
newlab0.shape
```

[14]:
```
(36143, 18)
```

[16]:
```
newlab0 = pd.DataFrame(newlab0)
```

[17]:
```
col_name = list(lab.columns.values)
newlab0.columns = col_name
```

[18]:
```
newlab02 = newlab0.iloc[36136:]
```

[19]:
```
newlab02['Encounter_ID'] = pd.DataFrame(newlab02['Encounter_ID']).
↪applymap(lambda x: x.replace('1e+05', '100000'))
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-
packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  """Entry point for launching an IPython kernel.
```

[20]:
```
newlab00 = newlab0.replace(newlab02)
```

[21]:
```
newlab00.describe()
```

[21]:

	CBC..ABSOLUTE.LYMPHOCYTES	CBC..ABSOLUTE.NEUTROPHILS	CBC..BASOPHILS	\
count	36143.000000	36143.000000	36143.000000	
mean	25.012467	70.014382	0.109371	
std	5.773137	5.765241	0.073498	
min	15.000000	60.000000	0.000000	
25%	20.000000	65.000000	0.100000	
50%	25.000000	70.000000	0.100000	
75%	30.000000	75.000000	0.200000	
max	35.000000	80.000000	0.200000	

	CBC..EOSINOPHILS	CBC..HEMATOCRIT	CBC..HEMOGLOBIN	\
count	36143.000000	36143.000000	36143.000000	
mean	0.340046	42.538489	14.508951	
std	0.154809	7.173616	2.598711	
min	0.100000	30.000000	10.000000	
25%	0.200000	36.400000	12.300000	
50%	0.300000	42.600000	14.500000	
75%	0.500000	48.700000	16.800000	
max	0.600000	55.000000	19.000000	

	CBC..PLATELET.COUNT	CBC..RED.BLOOD.CELL.COUNT	\
count	36143.000000	36143.000000	
mean	284.655720	4.998442	
std	95.354652	1.156924	
min	120.000000	3.000000	
25%	202.100000	4.000000	
50%	284.400000	5.000000	
75%	367.100000	6.000000	
max	450.000000	7.000000	

	CBC..WHITE.BLOOD.CELL.COUNT	METABOLIC..ALBUMIN	METABOLIC..BILI.TOTAL	\
count	36143.000000	36143.000000	36143.000000	

```
mean                    7.518798          4.249077                0.602128
std                     2.597528          1.013476                0.349107
min                     3.000000          2.500000                0.000000
25%                     5.200000          3.400000                0.300000
50%                     7.500000          4.200000                0.600000
75%                     9.800000          5.100000                0.900000
max                    12.000000          6.000000                1.200000

       METABOLIC..BUN  METABOLIC..CALCIUM  METABOLIC..CREATININE  \
count    36143.000000        36143.000000           36143.000000
mean        17.510342            9.501541               0.851078
std          7.183214            1.443756               0.206359
min          5.000000            7.000000               0.500000
25%         11.400000            8.300000               0.700000
50%         17.500000            9.500000               0.900000
75%         23.700000           10.700000               1.000000
max         30.000000           12.000000               1.200000

       METABOLIC..POTASSIUM  METABOLIC..SODIUM
count          36143.000000       36143.000000
mean               4.491819         140.038799
std                0.867464           8.668622
min                3.000000         125.000000
25%                3.700000         132.500000
50%                4.500000         140.100000
75%                5.200000         147.600000
max                6.000000         155.000000
```

5.0.3 Merge the three datasets into one correspondingly

[22]:
```
q2 = pd.merge(info, newlab00, on=['Encounter_ID'],how='left').merge(df,␣
↪on=['Encounter_ID'],how='left')
```

Drop redundant columns

[23]:
```
q2.drop(['Patient_ID', 'AdmissionStartDate', 'AdmissionEndDate', 'Lab_DTTM'],␣
↪axis=1, inplace=True)
```

[24]:
```
q2
```

[24]:
```
      Encounter_ID PatientGender       PatientRace  PatientEncounterAge  \
0         100000_2        Female             White            36.456455
1         100000_1        Female             White            21.408437
2         100000_5        Female             White            47.688073
3         100000_3        Female             White            42.671151
4         100000_7        Female             White            52.960276
5         100000_4        Female             White            47.277525
```

6	100000_6	Female	White	48.049906
7	100001_1	Female	African American	23.467454
8	100002_2	Female	Unknown	43.045790
9	100002_1	Female	Unknown	27.803538
10	100002_3	Female	Unknown	45.400816
11	100003_4	Female	White	88.133384
12	100003_3	Female	White	66.119312
13	100003_2	Female	White	65.950097
14	100003_1	Female	White	27.296089
15	100004_3	Female	African American	63.631083
16	100004_4	Female	African American	76.459379
17	100004_2	Female	African American	63.437677
18	100004_1	Female	African American	20.815506
19	100005_1	Male	Asian	19.801332
20	100005_5	Male	Asian	53.881459
21	100005_4	Male	Asian	50.096204
22	100005_2	Male	Asian	45.712839
23	100005_3	Male	Asian	46.095659
24	100006_2	Female	White	33.774769
25	100006_3	Female	White	56.929632
26	100006_5	Female	White	64.029558
27	100006_4	Female	White	61.401128
28	100006_1	Female	White	23.029740
29	100007_1	Female	White	27.463887
...	...	...	...	...
36113	109993_1	Male	White	25.429179
36114	109993_3	Male	White	36.122567
36115	109993_2	Male	White	30.011803
36116	109994_1	Male	White	21.940051
36117	109994_2	Male	White	28.596605
36118	109995_3	Female	African American	20.835092
36119	109995_1	Female	African American	19.565263
36120	109995_4	Female	African American	26.623121
36121	109995_2	Female	African American	20.085191
36122	109996_1	Male	White	24.693191
36123	109996_3	Male	White	56.546792
36124	109996_4	Male	White	83.947829
36125	109996_2	Male	White	34.139979
36126	109997_3	Male	White	36.050789
36127	109997_4	Male	White	37.225699
36128	109997_2	Male	White	21.418523
36129	109997_1	Male	White	19.429982
36130	109998_1	Male	White	19.880229
36131	109998_2	Male	White	40.956402
36132	109999_4	Female	Unknown	56.962831
36133	109999_2	Female	Unknown	41.844778
36134	109999_5	Female	Unknown	74.873390

36135	109999_6	Female	Unknown	79.156487
36136	109999_3	Female	Unknown	56.292777
36137	109999_1	Female	Unknown	18.771026
36138	110000_4	Male	African American	53.156375
36139	110000_3	Male	African American	46.765976
36140	110000_1	Male	African American	27.746113
36141	110000_2	Male	African American	30.322346
36142	110000_5	Male	African American	58.494166

	CBC..ABSOLUTE.LYMPHOCYTES	CBC..ABSOLUTE.NEUTROPHILS	CBC..BASOPHILS	\
0	16.1	77.9	0.2	
1	32.8	76.4	0.0	
2	27.4	62.2	0.1	
3	23.4	66.2	0.1	
4	30.3	64.4	0.1	
5	15.7	78.3	0.0	
6	27.1	66.1	0.1	
7	33.3	67.2	0.2	
8	17.2	70.7	0.0	
9	29.4	64.5	0.0	
10	34.7	74.8	0.2	
11	27.5	61.1	0.1	
12	22.4	77.5	0.1	
13	15.3	66.6	0.1	
14	29.5	75.8	0.1	
15	26.6	76.3	0.1	
16	25.4	79.4	0.1	
17	34.7	61.7	0.0	
18	29.9	70.2	0.0	
19	15.1	64.0	0.2	
20	23.4	65.5	0.2	
21	24.1	70.7	0.0	
22	16.0	60.3	0.2	
23	23.8	78.4	0.0	
24	22.5	61.2	0.1	
25	17.3	60.1	0.1	
26	15.9	71.4	0.2	
27	28.2	68.7	0.2	
28	25.3	64.1	0.2	
29	34.2	74.4	0.0	
...	...	...	...	
36113	31.1	62.6	0.2	
36114	18.8	76.6	0.0	
36115	28.5	78.1	0.2	
36116	28.0	73.8	0.0	
36117	33.8	60.0	0.1	
36118	22.4	74.9	0.0	

36119	33.0	77.8	0.0
36120	26.5	69.3	0.2
36121	33.1	70.7	0.1
36122	17.1	74.0	0.2
36123	28.3	61.6	0.1
36124	17.9	67.1	0.2
36125	22.2	69.1	0.0
36126	30.9	69.7	0.0
36127	16.8	64.1	0.0
36128	19.5	67.4	0.1
36129	31.6	72.8	0.2
36130	18.5	65.7	0.1
36131	28.0	74.0	0.2
36132	23.7	70.7	0.0
36133	22.9	78.8	0.2
36134	20.2	68.2	0.0
36135	28.4	71.4	0.1
36136	20.6	69.5	0.0
36137	30.3	75.5	0.0
36138	28.2	64.7	0.2
36139	27.0	78.5	0.1
36140	17.4	75.9	0.1
36141	19.9	60.1	0.1
36142	15.7	67.6	0.1

	CBC..EOSINOPHILS	CBC..HEMATOCRIT	CBC..HEMOGLOBIN	...	\
0	0.4	49.6	16.2	...	
1	0.1	35.9	14.8	...	
2	0.4	45.2	16.8	...	
3	0.4	40.4	18.7	...	
4	0.5	38.9	18.6	...	
5	0.2	54.4	18.9	...	
6	0.6	39.1	12.8	...	
7	0.4	40.2	18.1	...	
8	0.3	37.9	11.7	...	
9	0.3	42.9	13.4	...	
10	0.2	53.7	11.2	...	
11	0.4	38.8	11.6	...	
12	0.2	39.3	13.3	...	
13	0.4	53.3	13.8	...	
14	0.2	32.0	10.9	...	
15	0.3	35.1	15.1	...	
16	0.4	51.8	17.5	...	
17	0.5	48.1	14.2	...	
18	0.4	30.0	11.9	...	
19	0.5	41.3	13.3	...	
20	0.6	46.1	11.2	...	

```
21                0.5            47.1            15.5   ...
22                0.5            36.8            13.1   ...
23                0.1            31.6            12.9   ...
24                0.3            45.9            11.4   ...
25                0.2            39.2            10.1   ...
26                0.6            43.7            19.0   ...
27                0.4            33.4            14.8   ...
28                0.4            47.8            13.1   ...
29                0.5            44.4            18.9   ...
...               ...             ...             ...   ...
36113             0.2            46.7            10.3   ...
36114             0.5            35.5            16.3   ...
36115             0.2            45.1            14.7   ...
36116             0.1            41.1            17.7   ...
36117             0.5            38.0            14.6   ...
36118             0.6            40.5            12.1   ...
36119             0.4            53.8            15.1   ...
36120             0.4            33.0            18.4   ...
36121             0.4            35.0            10.2   ...
36122             0.5            43.5            14.1   ...
36123             0.3            46.3            16.9   ...
36124             0.6            53.6            12.0   ...
36125             0.3            42.0            15.8   ...
36126             0.3            51.8            13.2   ...
36127             0.6            41.4            17.1   ...
36128             0.1            55.0            10.7   ...
36129             0.2            50.1            14.7   ...
36130             0.6            36.3            15.4   ...
36131             0.4            45.8            14.0   ...
36132             0.6            30.7            15.0   ...
36133             0.2            54.8            18.4   ...
36134             0.5            38.5            18.9   ...
36135             0.5            41.5            17.5   ...
36136             0.5            44.7            17.3   ...
36137             0.5            37.2            17.5   ...
36138             0.2            38.9            13.6   ...
36139             0.6            46.5            17.1   ...
36140             0.1            41.0            14.1   ...
36141             0.6            32.3            15.3   ...
36142             0.6            49.7            14.5   ...

       CBC..RED.BLOOD.CELL.COUNT  CBC..WHITE.BLOOD.CELL.COUNT  \
0                            5.8                          6.4
1                            4.8                         10.7
2                            4.9                         11.8
3                            3.2                          8.5
4                            5.3                          9.1
```

5	5.9	7.0
6	6.8	8.0
7	6.6	4.7
8	3.3	9.2
9	5.9	6.4
10	6.9	10.6
11	6.6	4.8
12	5.3	3.0
13	3.7	11.9
14	3.1	3.7
15	5.0	11.7
16	5.6	7.5
17	3.7	9.2
18	4.0	10.3
19	5.4	10.6
20	3.7	4.1
21	5.5	5.7
22	4.2	11.7
23	6.5	6.0
24	4.4	8.7
25	3.9	7.0
26	6.2	3.5
27	5.6	4.7
28	3.2	4.1
29	5.5	8.0
...	...	...
36113	4.5	10.5
36114	6.4	3.1
36115	6.4	6.8
36116	4.9	7.1
36117	3.1	7.4
36118	5.3	8.3
36119	6.7	9.8
36120	4.9	3.3
36121	6.3	10.8
36122	3.2	3.3
36123	6.2	5.5
36124	3.5	9.4
36125	5.6	11.2
36126	3.1	11.4
36127	4.6	3.2
36128	4.9	11.8
36129	6.3	6.2
36130	3.3	9.9
36131	4.9	3.8
36132	5.9	8.9
36133	5.9	5.0

36134	6.8	4.4
36135	7.0	4.2
36136	5.4	5.1
36137	3.4	4.6
36138	6.6	6.4
36139	3.2	9.8
36140	6.4	9.7
36141	3.9	9.6
36142	6.4	6.4

	METABOLIC..ALBUMIN	METABOLIC..BILI.TOTAL	METABOLIC..BUN	\
0	2.7	0.9	15.6	
1	5.8	0.8	12.6	
2	5.9	0.1	21.4	
3	3.7	0.3	15.1	
4	4.1	1.2	5.2	
5	4.0	0.9	24.4	
6	4.8	1.1	27.7	
7	3.5	1.2	20.5	
8	3.4	0.7	14.7	
9	3.2	0.0	25.9	
10	4.1	0.7	7.1	
11	5.5	0.3	10.2	
12	5.7	1.1	25.9	
13	4.2	0.4	15.5	
14	4.1	0.2	26.6	
15	3.6	0.1	25.2	
16	5.4	0.9	27.9	
17	2.5	1.0	23.1	
18	2.7	0.2	8.8	
19	2.6	0.6	9.9	
20	5.5	0.9	13.8	
21	5.5	0.7	29.8	
22	3.1	0.8	16.5	
23	2.5	1.0	23.7	
24	4.0	1.2	15.2	
25	3.2	0.3	16.9	
26	4.0	0.1	29.7	
27	5.6	0.6	24.8	
28	2.7	0.8	20.6	
29	2.8	0.7	26.1	
...	...	...	...	
36113	4.6	0.8	28.3	
36114	5.7	0.3	11.6	
36115	2.5	0.6	5.5	
36116	3.4	1.1	15.9	
36117	3.9	0.0	26.7	

36118	3.1	0.3	26.6
36119	3.2	1.1	11.4
36120	5.3	0.5	13.6
36121	4.0	0.9	18.6
36122	4.9	0.5	27.0
36123	3.0	0.9	7.5
36124	4.9	0.3	11.9
36125	4.5	0.6	25.0
36126	3.6	1.1	6.2
36127	5.6	0.4	24.4
36128	3.4	0.2	23.1
36129	2.8	0.5	10.5
36130	4.4	1.0	16.3
36131	3.7	0.4	22.5
36132	4.4	0.0	25.0
36133	4.7	0.4	24.4
36134	3.3	1.1	23.7
36135	5.7	0.6	20.5
36136	3.0	0.3	9.9
36137	3.2	0.5	22.1
36138	5.7	1.1	23.0
36139	3.2	0.4	5.4
36140	2.8	0.4	8.7
36141	4.5	1.1	25.5
36142	3.8	0.1	9.0

	METABOLIC..CALCIUM	METABOLIC..CREATININE	METABOLIC..POTASSIUM	\
0	8.1	1.2	4.7	
1	11.1	0.9	4.1	
2	7.5	0.8	5.6	
3	7.9	0.7	5.3	
4	7.2	1.2	5.7	
5	7.0	1.1	4.4	
6	8.4	0.7	3.4	
7	11.3	1.1	4.4	
8	11.9	1.2	5.7	
9	11.8	0.7	3.3	
10	12.0	1.2	5.8	
11	8.3	0.7	5.3	
12	11.1	1.2	4.7	
13	8.4	1.2	3.1	
14	9.6	0.9	5.0	
15	10.1	0.5	3.8	
16	10.0	0.5	5.5	
17	9.3	0.9	4.3	
18	7.4	1.0	4.0	
19	7.4	1.0	3.8	

20	9.6	0.7	4.8
21	8.8	0.7	4.0
22	10.4	0.7	5.1
23	8.7	0.6	5.7
24	8.1	0.8	4.7
25	8.5	1.2	5.1
26	7.5	0.9	3.5
27	7.8	1.0	5.5
28	10.8	0.5	3.3
29	11.9	1.0	3.3
...	...	...	...
36113	7.4	0.5	4.4
36114	8.3	0.7	5.8
36115	11.1	1.1	6.0
36116	8.2	0.6	4.4
36117	11.8	0.7	3.4
36118	9.8	1.1	5.6
36119	8.9	1.0	3.2
36120	11.2	1.0	3.5
36121	10.7	0.7	5.6
36122	8.1	0.9	4.7
36123	9.5	0.9	5.2
36124	9.7	0.6	5.2
36125	8.6	0.9	5.0
36126	9.5	1.0	5.1
36127	11.2	1.1	3.2
36128	7.1	0.9	5.1
36129	7.0	1.0	6.0
36130	7.8	1.0	5.8
36131	11.8	0.6	5.8
36132	11.6	0.5	4.0
36133	9.1	1.2	5.8
36134	11.2	0.9	4.7
36135	8.3	0.7	3.7
36136	8.3	1.1	3.8
36137	11.8	0.8	5.0
36138	10.9	0.7	4.1
36139	11.6	0.7	4.7
36140	7.2	0.7	3.8
36141	11.3	0.6	3.1
36142	9.0	0.6	3.5

	METABOLIC..SODIUM	outcome
0	140.9	0
1	135.5	0
2	137.1	0
3	150.0	0

4	151.7	0
5	155.0	0
6	139.5	0
7	137.6	0
8	129.1	0
9	145.2	0
10	147.6	0
11	126.6	0
12	136.0	0
13	144.1	0
14	125.1	0
15	128.8	0
16	140.4	0
17	133.2	0
18	148.0	0
19	145.9	0
20	150.7	0
21	127.1	0
22	146.7	0
23	126.1	0
24	138.2	0
25	134.5	0
26	145.9	0
27	142.9	0
28	149.5	0
29	131.3	0
...	...	...
36113	133.3	0
36114	129.8	0
36115	127.9	0
36116	149.8	0
36117	130.4	0
36118	130.8	0
36119	132.4	0
36120	139.6	0
36121	136.6	0
36122	153.6	0
36123	134.8	0
36124	134.2	0
36125	128.0	0
36126	149.9	0
36127	154.1	0
36128	132.4	0
36129	152.9	0
36130	145.1	0
36131	139.6	0
36132	145.9	0

```
36133              139.8        0
36134              129.3        0
36135              133.2        0
36136              154.4        0
36137              150.7        0
36138              130.8        0
36139              143.0        0
36140              137.4        0
36141              140.8        0
36142              131.1        0

[36143 rows x 21 columns]
```

[25]:
```
q2.describe()
```

[25]:
```
       PatientEncounterAge  CBC..ABSOLUTE.LYMPHOCYTES  \
count         36143.000000               36143.000000
mean             41.748503                  25.012467
std              18.058317                   5.773137
min              18.011880                  15.000000
25%              25.839463                  20.000000
50%              38.588578                  25.000000
75%              54.398018                  30.000000
max              92.958042                  35.000000

       CBC..ABSOLUTE.NEUTROPHILS  CBC..BASOPHILS  CBC..EOSINOPHILS  \
count                36143.000000    36143.000000      36143.000000
mean                    70.014382        0.109371          0.340046
std                      5.765241        0.073498          0.154809
min                     60.000000        0.000000          0.100000
25%                     65.000000        0.100000          0.200000
50%                     70.000000        0.100000          0.300000
75%                     75.000000        0.200000          0.500000
max                     80.000000        0.200000          0.600000

       CBC..HEMATOCRIT  CBC..HEMOGLOBIN  CBC..PLATELET.COUNT  \
count     36143.000000     36143.000000         36143.000000
mean         42.538489        14.508951           284.655720
std           7.173616         2.598711            95.354652
min          30.000000        10.000000           120.000000
25%          36.400000        12.300000           202.100000
50%          42.600000        14.500000           284.400000
75%          48.700000        16.800000           367.100000
max          55.000000        19.000000           450.000000

       CBC..RED.BLOOD.CELL.COUNT  CBC..WHITE.BLOOD.CELL.COUNT  \
count                36143.000000                 36143.000000
mean                     4.998442                     7.518798
```

```
std                    1.156924                    2.597528
min                    3.000000                    3.000000
25%                    4.000000                    5.200000
50%                    5.000000                    7.500000
75%                    6.000000                    9.800000
max                    7.000000                   12.000000

       METABOLIC..ALBUMIN  METABOLIC..BILI.TOTAL  METABOLIC..BUN  \
count        36143.000000           36143.000000    36143.000000
mean             4.249077               0.602128       17.510342
std              1.013476               0.349107        7.183214
min              2.500000               0.000000        5.000000
25%              3.400000               0.300000       11.400000
50%              4.200000               0.600000       17.500000
75%              5.100000               0.900000       23.700000
max              6.000000               1.200000       30.000000

       METABOLIC..CALCIUM  METABOLIC..CREATININE  METABOLIC..POTASSIUM  \
count        36143.000000           36143.000000          36143.000000
mean             9.501541               0.851078              4.491819
std              1.443756               0.206359              0.867464
min              7.000000               0.500000              3.000000
25%              8.300000               0.700000              3.700000
50%              9.500000               0.900000              4.500000
75%             10.700000               1.000000              5.200000
max             12.000000               1.200000              6.000000

       METABOLIC..SODIUM       outcome
count       36143.000000  36143.000000
mean          140.038799      0.003541
std             8.668622      0.059406
min           125.000000      0.000000
25%           132.500000      0.000000
50%           140.100000      0.000000
75%           147.600000      0.000000
max           155.000000      1.000000
```

[26]:
```
q2.dtypes
```

[26]:
```
Encounter_ID                  object
PatientGender                 object
PatientRace                   object
PatientEncounterAge          float64
CBC..ABSOLUTE.LYMPHOCYTES    float64
CBC..ABSOLUTE.NEUTROPHILS    float64
CBC..BASOPHILS               float64
CBC..EOSINOPHILS             float64
CBC..HEMATOCRIT              float64
```

```
CBC..HEMOGLOBIN                float64
CBC..PLATELET.COUNT            float64
CBC..RED.BLOOD.CELL.COUNT      float64
CBC..WHITE.BLOOD.CELL.COUNT    float64
METABOLIC..ALBUMIN             float64
METABOLIC..BILI.TOTAL          float64
METABOLIC..BUN                 float64
METABOLIC..CALCIUM             float64
METABOLIC..CREATININE          float64
METABOLIC..POTASSIUM           float64
METABOLIC..SODIUM              float64
outcome                          int64
dtype: object
```

6 Section 1, (Q3):

[27]:
```
newlab03 = newlab00.copy()
```

[28]:
```
newlab03['mean_labs']= newlab03.iloc[:, -16:].sum(axis=1)/16
```

[29]:
```
newlab03 = newlab03[['Encounter_ID','mean_labs']]
```

[30]:
```
q3 = pd.merge(info, newlab03, on=['Encounter_ID'],how='left').merge(df,␣
 ↪on=['Encounter_ID'],how='left')
q3.drop(['Patient_ID', 'AdmissionStartDate', 'AdmissionEndDate'], axis=1,␣
 ↪inplace=True)
q3.describe()
```

[30]:

	PatientEncounterAge	mean_labs	outcome
count	36143.000000	36143.000000	36143.000000
mean	41.748503	39.183841	0.003541
std	18.058317	6.047758	0.059406
min	18.011880	26.262500	0.000000
25%	25.839463	34.009375	0.000000
50%	38.588578	39.143750	0.000000
75%	54.398018	44.356250	0.000000
max	92.958042	52.781250	1.000000

[31]:
```
q3.head()
```

[31]:

	Encounter_ID	PatientGender	PatientRace	PatientEncounterAge	mean_labs	outcome
0	100000_2	Female	White	36.456455	42.86875	
1	100000_1	Female	White	21.408437	45.76875	
2	100000_5	Female	White	47.688073	41.41250	
3	100000_3	Female	White	42.671151	47.03750	
4	100000_7	Female	White	52.960276	43.47500	

```
0       0
1       0
2       0
3       0
4       0
```

[32]:
```
q3.dtypes
```

[32]:
```
Encounter_ID            object
PatientGender           object
PatientRace             object
PatientEncounterAge    float64
mean_labs              float64
outcome                  int64
dtype: object
```

7 Section 1, (Q4):

[33]:
```
df30.head()
```

[33]:
```
      Patient_ID Encounter_ID       AdmissionStartDate  \
373       100103     100103_1  1982-06-29 21:16:48.833
433       100118     100118_5  2005-06-15 14:36:49.753
474       100131     100131_1  1980-10-06 00:29:48.673
637       100174     100174_1  1991-09-11 18:28:18.557
1169      100331     100331_6  2007-12-07 23:28:17.110

             AdmissionEndDate PatientGender       PatientRace  \
373   1982-07-12 01:09:45.967          Male             Asian
433   2005-06-29 12:55:34.900          Male             White
474   1980-10-11 08:05:28.010        Female             White
637   1991-09-24 20:24:47.230          Male  African American
1169  2007-12-16 16:44:45.767          Male           Unknown

      PatientEncounterAge  outcome
373             18.133936        1
433             49.796189        1
474             26.301427        1
637             25.933616        1
1169            75.158844        1
```

[34]:
```
df30s04 = df30[df30["AdmissionEndDate"].str[:4].astype(int)<=2004]
```

[35]:
```
df30g04 = df30[df30["AdmissionEndDate"].str[:4].astype(int)>2004]
```

[36]:
```
# Patients who were re-admitted within 30 days and years <= 2004
len(df30s04.index)
```

[36]:
```
78
```

[37]:
```
# Patients who were re-admitted within 30 days and years > 2004
len(df30g04.index)
```

[37]: 50

[38]:
```
# Patients who were not re-admitted within 30 days and years <= 2004
dfnot30s04 = dfnot30[dfnot30["AdmissionEndDate"].str[:4].astype(int)<=2004]
dfnot30g04 = dfnot30[dfnot30["AdmissionEndDate"].str[:4].astype(int)>2004]
len(dfnot30s04.index)
```

[38]: 21416

[39]:
```
# Patients who were not re-admitted within 30 days and years > 2004
len(dfnot30g04.index)
```

[39]: 14599

Total for the training with percentage

[40]:
```
len(df30s04.index)+len(dfnot30s04.index)
```

[40]: 21494

[41]:
```
print(len(df30s04.index)/(len(df30s04.index)+len(dfnot30s04.index))*100,"%")
```

0.36289196985205174 %

[42]:
```
print(len(dfnot30s04.index)/(len(df30s04.index)+len(dfnot30s04.index))*100,"%")
```

99.63710803014794 %

Total for the test set with percentage

[43]:
```
len(df30g04.index)+len(dfnot30g04.index)
```

[43]: 14649

[44]:
```
print(len(df30g04.index)/(len(df30g04.index)+len(dfnot30g04.index))*100,"%")
```

0.34132022663663053 %

[45]:
```
print(len(dfnot30g04.index)/(len(df30g04.index)+len(dfnot30g04.index))*100,"%")
```

99.65867977336337 %

---

8 Section 2: GUSTO 30-day Mortality Prediction

8.1 Data preparation

[46]:
```
gusto = pd.read_csv("gusto_data.csv")
gusto.head()
```

[46]:

	DAY30	AGE	A65	SEX	KILLIP	SHO	DIA	HYP	HRT	ANT	...	WEI	\
0	0	70.313	1	0	1	0	0	0	0	1	...	84.0	
1	0	59.844	0	0	1	0	1	0	0	1	...	115.0	
2	0	59.023	0	0	1	0	0	0	1	0	...	76.0	
3	1	80.375	1	1	1	0	0	0	1	0	...	50.0	
4	0	64.750	0	0	1	0	0	1	0	0	...	97.4	

	SMK	HTN	LIP	PAN	FAM	STE	ST4	TTR	GROUP
0	3	1	1	0	0	1	0	1	west
1	1	1	0	0	1	6	1	0	west
2	1	1	0	0	1	3	0	0	west
3	3	0	0	0	0	3	0	0	west
4	1	0	0	1	1	2	0	1	west

[5 rows x 23 columns]

[47]:
```
gusto.shape
```

[47]:
```
(3661, 23)
```

[48]:
```
gusto0=gusto.copy()
gusto0.dropna()
gusto0.shape
```

[48]:
```
(3661, 23)
```

9 Section 2, (Q1):

9.0.1 Total number

[49]:
```
columns2 = ['AGE','SEX','GROUP','DAY30']
categorical2 = ['SEX','GROUP']
groupby2 = 'DAY30'
mytable2 = TableOne(gusto, columns2, categorical2, groupby2, pval=True)
print(mytable2)
```

```
                 Grouped by DAY30
                          isnull            0            1    pval
ptest
variable level
n                                         3430          231
AGE                            0  60.2 (11.6)  71.1 (10.7)  <0.001  Two Sample
T-test
SEX      0                     0  2581 (75.2)   140 (60.6)  <0.001        Chi-
squared
         1                        849 (24.8)    91 (39.4)
```

```
GROUP    sample2                0    239 (7.0)     20 (8.7)   0.694        Chi-
squared
         sample4                   733 (21.4)    52 (22.5)
         sample5                   405 (11.8)    24 (10.4)
         west                     2053 (59.9)   135 (58.4)
[1] Warning, test for normality reports non-normal distributions for: AGE.
```

[50]:
```
g30 = gusto[gusto['DAY30']==1]
gnot30 = gusto[gusto['DAY30']==0]
ttest_ind(g30['AGE'], gnot30['AGE'])
```

[50]:
```
Ttest_indResult(statistic=13.957560682198121, pvalue=3.5536013952145326e-43)
```

10 Section 2, (Q2):

Patients who died within 30 days

[51]:
```
len(g30.loc[(g30['GROUP'] == 'sample2'
            ) | (g30['GROUP'] == 'sample4'
                ) | (g30['GROUP'] == 'sample5')].index)
```

[51]:
```
96
```

[52]:
```
len(g30[g30['GROUP']=='west'].index)
```

[52]:
```
135
```

Patients who were alive at 30 days

[53]:
```
len(gnot30.loc[(gnot30['GROUP'] == 'sample2'
               ) | (gnot30['GROUP'] == 'sample4'
                   ) | (gnot30['GROUP'] == 'sample5')].index)
```

[53]:
```
1377
```

[54]:
```
len(gnot30[gnot30['GROUP']=='west'].index)
```

[54]:
```
2053
```

Total and percentage for training

[55]:
```
len(g30.loc[(g30['GROUP'] == 'sample2'
            ) | (g30['GROUP'] == 'sample4'
                ) | (g30['GROUP'] == 'sample5')].index)+len(gnot30.
 ↪loc[(gnot30['GROUP'] == 'sample2'
               ) | (gnot30['GROUP'] == 'sample4'
                   ) | (gnot30['GROUP'] == 'sample5')].index)
```

[55]:
```
1473
```

[56]:
```
print(len(g30.loc[(g30['GROUP'] == 'sample2'
                  ) | (g30['GROUP'] == 'sample4'
                      ) | (g30['GROUP'] == 'sample5'
                          )].index)/(len(g30.loc[(g30['GROUP'] == 'sample2'
                                                ) | (g30['GROUP'] == 'sample4'
                                                    ) | (g30['GROUP'] ==␣
↪'sample5'
                                                        )].index)+len(
   gnot30.loc[(gnot30['GROUP'] == 'sample2'
               ) | (gnot30['GROUP'] == 'sample4'
                   ) | (gnot30['GROUP'] == 'sample5'
                       )].index))*100,"%")
```

```
6.517311608961303 %
```

[57]:
```
print(len(gnot30.loc[(gnot30['GROUP'] == 'sample2'
                     ) | (gnot30['GROUP'] == 'sample4'
                         ) | (gnot30['GROUP'] == 'sample5'
                             )].index)/(len(g30.loc[(g30['GROUP'] == 'sample2'
                                                   ) | (g30['GROUP'] ==␣
↪'sample4'
                                                       ) | (g30['GROUP'] ==␣
↪'sample5'
                                                           )].index)+len(
   gnot30.loc[(gnot30['GROUP'] == 'sample2'
               ) | (gnot30['GROUP'] == 'sample4'
                   ) | (gnot30['GROUP'] == 'sample5'
                       )].index))*100,"%")
```

```
93.48268839103869 %
```

Total and percentage for test set

[58]:
```
len(g30[g30['GROUP']=='west'].index)+len(gnot30[gnot30['GROUP']=='west'].index)
```

[58]:
```
2188
```

[59]:
```
print(len(g30[g30['GROUP']=='west'
              ].index)/(len(g30[g30['GROUP']=='west'
                                ].index)+len(gnot30[gnot30['GROUP'
                                                           ]=='west'].
↪index))*100,"%")
```

```
6.170018281535649 %
```

[60]:
```
print(len(gnot30[gnot30['GROUP']=='west'
                 ].index)/(len(g30[g30['GROUP']=='west'
                                   ].index)+len(gnot30[gnot30['GROUP'
```

```
                                                                ]=='west'].
↪index))*100,"%")
```

```
93.82998171846435 %
```

11 Section 2, (Q3):

[61]:
```
g30_train = g30.loc[(g30['GROUP'] == 'sample2'
                    ) | (g30['GROUP'] == 'sample4'
                        ) | (g30['GROUP'] == 'sample5')]
gnot30_train = gnot30.loc[(gnot30['GROUP'] == 'sample2'
                           ) | (gnot30['GROUP'] == 'sample4'
                               ) | (gnot30['GROUP'] == 'sample5')]
gusto_train = pd.concat([g30_train,gnot30_train])
gusto_train = gusto_train.reset_index(drop=True)
gusto_train.describe()
```

[61]:

	DAY30	AGE	A65	SEX	KILLIP	\
count	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	
mean	0.065173	61.415623	0.410726	0.268839	1.194840	
std	0.246915	11.448781	0.492133	0.443507	0.462655	
min	0.000000	25.891000	0.000000	0.000000	1.000000	
25%	0.000000	52.578000	0.000000	0.000000	1.000000	
50%	0.000000	62.242000	0.000000	0.000000	1.000000	
75%	0.000000	70.469000	1.000000	1.000000	1.000000	
max	1.000000	88.828000	1.000000	1.000000	4.000000	

	SHO	DIA	HYP	HRT	ANT	\
count	1473.000000	1473.000000	1473.000000	1473.000000	1473.000000	
mean	0.020367	0.114732	0.073320	0.292600	0.361168	
std	0.141299	0.318806	0.260749	0.455111	0.480502	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	1.000000	1.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	

	...	HEI	WEI	SMK	HTN	\
count	...	1473.000000	1473.000000	1473.000000	1473.000000	
mean	...	170.338900	78.200068	1.909029	0.385608	
std	...	9.779777	16.531963	0.803275	0.486904	
min	...	141.000000	37.000000	1.000000	0.000000	
25%	...	163.800000	68.000000	1.000000	0.000000	
50%	...	170.200000	77.000000	2.000000	0.000000	
75%	...	177.300000	87.000000	3.000000	1.000000	
max	...	199.700000	180.000000	3.000000	1.000000	

```
                LIP          PAN          FAM          STE          ST4  \
count  1473.000000  1473.000000  1473.000000  1473.000000  1473.000000
mean      0.386286     0.364562     0.428377     4.150034     0.386965
std       0.487063     0.481471     0.495012     1.865345     0.487221
min       0.000000     0.000000     0.000000     0.000000     0.000000
25%       0.000000     0.000000     0.000000     3.000000     0.000000
50%       0.000000     0.000000     0.000000     4.000000     0.000000
75%       1.000000     1.000000     1.000000     6.000000     1.000000
max       1.000000     1.000000     1.000000    10.000000     1.000000

                TTR
count  1473.000000
mean      0.560760
std       0.496463
min       0.000000
25%       0.000000
50%       1.000000
75%       1.000000
max       1.000000

[8 rows x 22 columns]
```

[62]:
```
gusto_train.dtypes
```

[62]:
```
DAY30       int64
AGE       float64
A65         int64
SEX         int64
KILLIP      int64
SHO         int64
DIA         int64
HYP         int64
HRT         int64
ANT         int64
PMI         int64
HIG         int64
HEI       float64
WEI       float64
SMK         int64
HTN         int64
LIP         int64
PAN         int64
FAM         int64
STE         int64
ST4         int64
TTR         int64
GROUP      object
dtype: object
```

[63]:
```
g30_test = g30[g30['GROUP']=='west']
gnot30_test = gnot30[gnot30['GROUP']=='west']
gusto_test = pd.concat([g30_test,gnot30_test])
gusto_test = gusto_test.reset_index(drop=True)
gusto_test.describe()
```

[63]:

	DAY30	AGE	A65	SEX	KILLIP	\
count	2188.000000	2188.000000	2188.000000	2188.000000	2188.000000	
mean	0.061700	60.469186	0.383455	0.248629	1.132084	
std	0.240665	12.026568	0.486339	0.432317	0.409550	
min	0.000000	23.910000	0.000000	0.000000	1.000000	
25%	0.000000	50.932000	0.000000	0.000000	1.000000	
50%	0.000000	60.547000	0.000000	0.000000	1.000000	
75%	0.000000	69.922000	1.000000	0.000000	1.000000	
max	1.000000	89.484000	1.000000	1.000000	4.000000	

	SHO	DIA	HYP	HRT	ANT	\
count	2188.000000	2188.000000	2188.000000	2188.000000	2188.000000	
mean	0.014625	0.142596	0.096435	0.333638	0.372486	
std	0.120075	0.349740	0.295254	0.471620	0.483577	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	1.000000	1.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	

	...	HEI	WEI	SMK	HTN	\
count	...	2188.000000	2188.000000	2188.000000	2188.000000	
mean	...	172.129936	82.888940	1.866545	0.403565	
std	...	10.094343	17.692498	0.821252	0.490724	
min	...	140.900000	36.000000	1.000000	0.000000	
25%	...	165.100000	70.900000	1.000000	0.000000	
50%	...	173.000000	82.000000	2.000000	0.000000	
75%	...	180.000000	92.050000	3.000000	1.000000	
max	...	205.700000	180.000000	3.000000	1.000000	

	LIP	PAN	FAM	STE	ST4	\
count	2188.000000	2188.000000	2188.000000	2188.000000	2188.000000	
mean	0.404936	0.340494	0.475777	3.999543	0.356033	
std	0.490992	0.473984	0.499527	1.878451	0.478935	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	3.000000	0.000000	
50%	0.000000	0.000000	0.000000	3.000000	0.000000	
75%	1.000000	1.000000	1.000000	5.000000	1.000000	
max	1.000000	1.000000	1.000000	11.000000	1.000000	

	TTR

```
count  2188.000000
mean      0.608775
std       0.488136
min       0.000000
25%       0.000000
50%       1.000000
75%       1.000000
max       1.000000

[8 rows x 22 columns]
```

[64]:
```
gusto_test.dtypes
```

[64]:
```
DAY30       int64
AGE       float64
A65         int64
SEX         int64
KILLIP      int64
SHO         int64
DIA         int64
HYP         int64
HRT         int64
ANT         int64
PMI         int64
HIG         int64
HEI       float64
WEI       float64
SMK         int64
HTN         int64
LIP         int64
PAN         int64
FAM         int64
STE         int64
ST4         int64
TTR         int64
GROUP      object
dtype: object
```

[ ]:

[ ]:

[ ]: