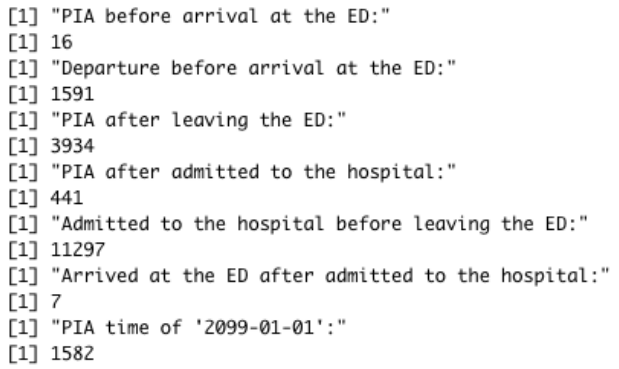
STA2453 Lab 1 Report - Yihan Duan (1003118547)

INTRODUCTION

This report focuses on: data quality issues their (potential) fixes, descriptive analysis on LOS, time to PIA, complaints, CTAS and number of encounters per day, difference in the average volume between weekends and weekdays, and analysis on census by time of day.

DATA QUALITY ISSUES

1) **Duplicate records** are found on **date (ed_start_time) 2019-07-18** from ENCOUNTER_NUM 43934 to 44149. **215 duplicate entries** removed. Maybe we should remove these entries from the database as well?
2) **Missing values** are found for column **ed_start_time** and **ed_end_time**. There are **793** entries missing ed_start_time, **396** missing ed_end_time and 3 missing both. Analysis that involves either column are done with these rows removed. Should we fill in an estimated time based on average time? Is there a way to retrieve these data?
3) There are **371 records that are admitted** but have **no adm_start_time**.
4) There are 2 entries at **ENCOUNTER_NUM 44042** that has **different ed_pia_time**, **11:02** and **07:02** respectively. Which one should I keep?
5) **Unrealistic order** of arrival, PIA, departure and admission time:

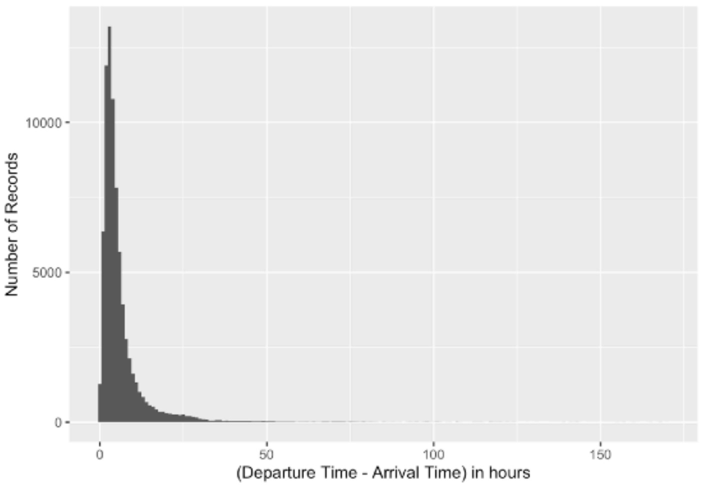


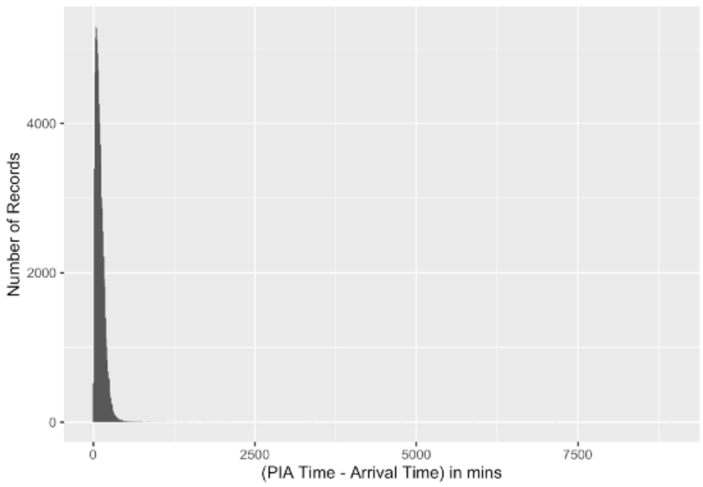
```
[1] "PIA before arrival at the ED:"
[1] 16
[1] "Departure before arrival at the ED:"
[1] 1591
[1] "PIA after leaving the ED:"
[1] 3934
[1] "PIA after admitted to the hospital:"
[1] 441
[1] "Admitted to the hospital before leaving the ED:"
[1] 11297
[1] "Arrived at the ED after admitted to the hospital:"
[1] 7
[1] "PIA time of '2099-01-01':"
[1] 1582
```

Analysis involving time calculation ignores these records records. Is this okay?

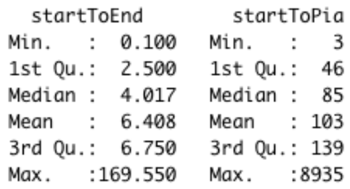
6) **35** records on **date (ed_start_time) 2020-01-01**. I kept these records. Should I remove?
7) **Los have upper bound of 24**. We ignore los variable in calculating the actual los.
8) Transformation mentioned in class. Is it okay?

DESCRIPTIVE ANALYSIS



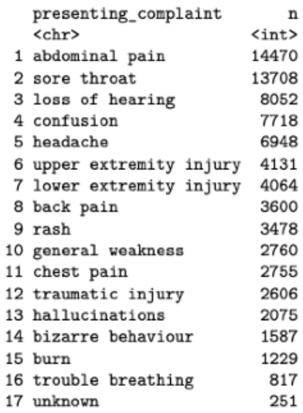


1) **Length of Stay (in hours)** have **mean 6.417152** and **standard deviation 9.136179**. The distribution of length of stay is right skewed. We can probably model it with gamma or log-normal distribution.
   Note that all negative values are ignored.
2) **Time to PIA (in mins)** have **mean 102.8916** and **standard deviation 104.6582**. The distribution is also right skewed.
   Note that all negative values and all PIA time on 2099 are ignored.

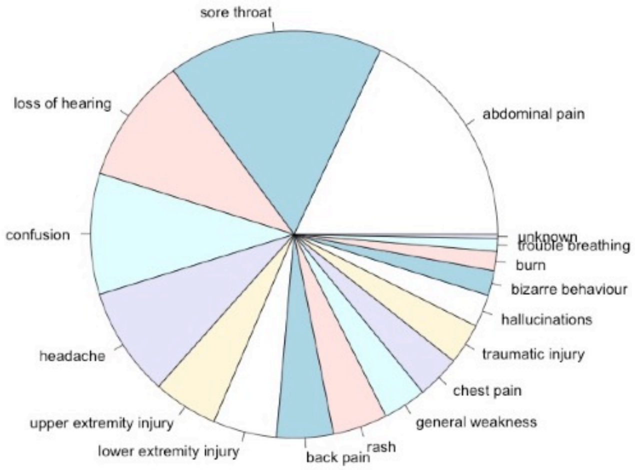


```
  startToEnd       startToPia
Min.   :  0.100   Min.   :   3
1st Qu.:  2.500   1st Qu.:  46
Median :  4.017   Median :  85
Mean   :  6.408   Mean   : 103
3rd Qu.:  6.750   3rd Qu.: 139
Max.   :169.550   Max.   :8935
```

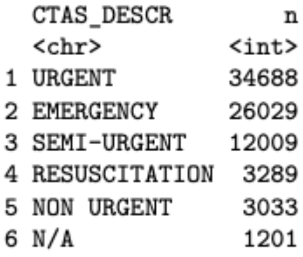
3) **Counts and proportions of presenting complaints**



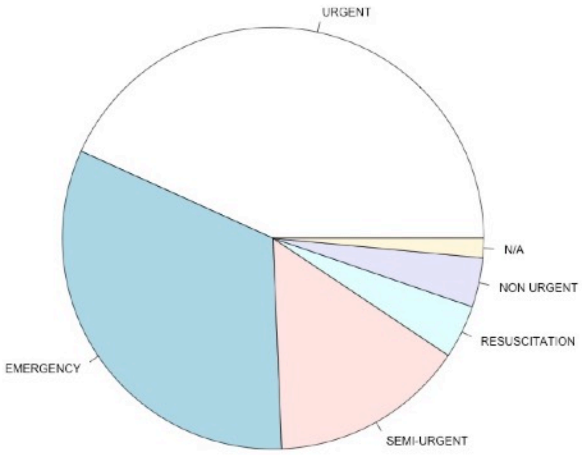
```
   presenting_complaint       n
   <chr>                  <int>
 1 abdominal pain         14470
 2 sore throat            13708
 3 loss of hearing         8052
 4 confusion               7718
 5 headache                6948
 6 upper extremity injury  4131
 7 lower extremity injury  4064
 8 back pain               3600
 9 rash                    3478
10 general weakness        2760
11 chest pain              2755
12 traumatic injury        2606
13 hallucinations          2075
14 bizarre behaviour       1587
15 burn                    1229
16 trouble breathing        817
17 unknown                  251
```



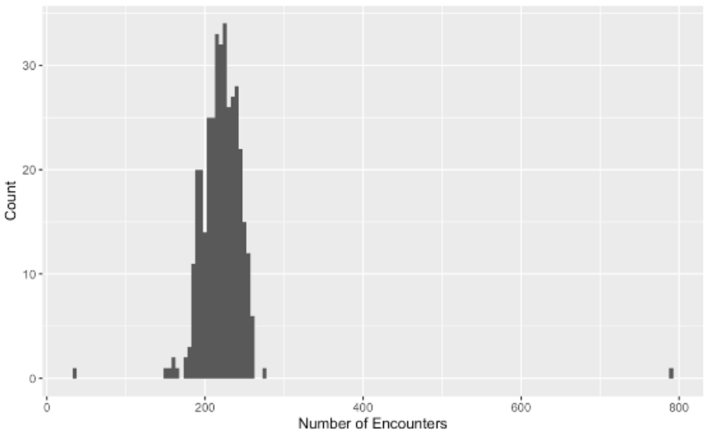
4) **Counts and proportions of CTAS**

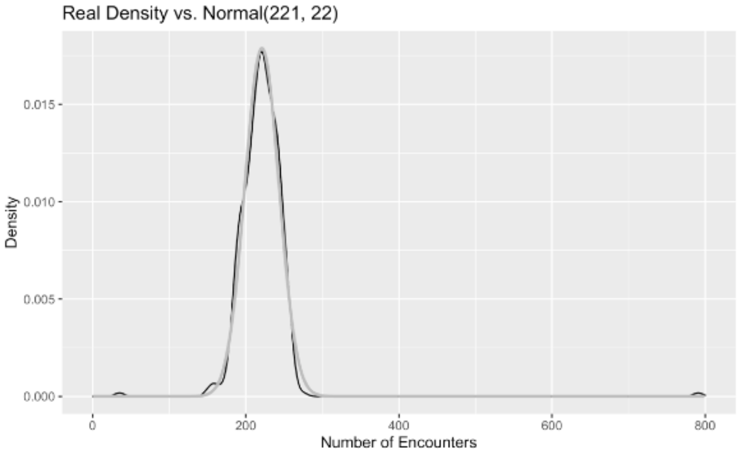


```
  CTAS_DESCR         n
  <chr>          <int>
1 URGENT         34688
2 EMERGENCY      26029
3 SEMI-URGENT    12009
4 RESUSCITATION   3289
5 NON URGENT      3033
6 N/A             1201
```

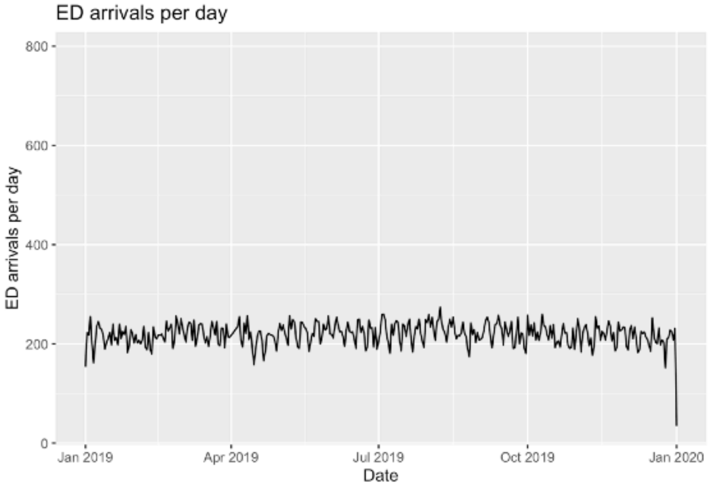


5) **# of Encounters Per day**



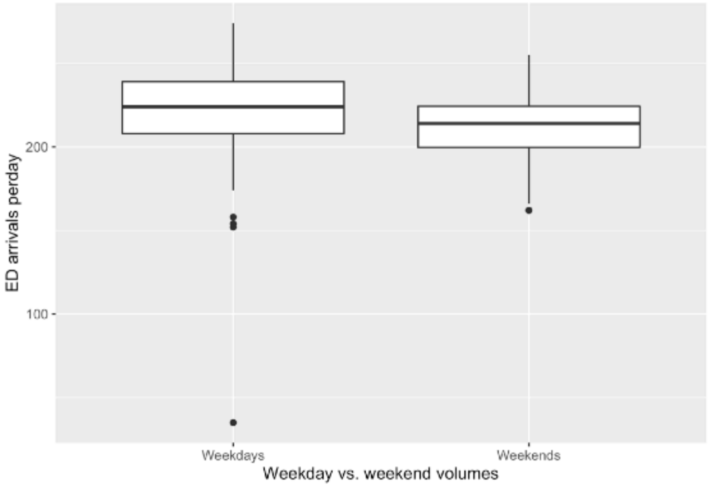


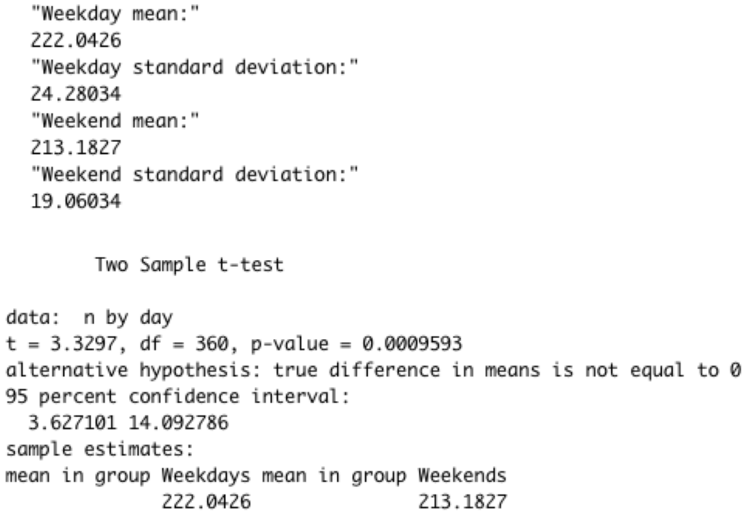
Mean: 221.0716. Standard deviation: 37.91762. As we ca see, apart from two extreme values, the number of encounters per day loosely follows a Norm(221, 22).



Looking at the time series plot, we can see that the number of arrivals is close the **mean value of 221** throughout the year. Note that this is the data after we removed all duplicate entries. So the spike in the assignment description is not here. However, we are still concerned about the data on ’01-01-2020’, as only 35 records exists for this day. Comparing to all the other existing volumes, this is very low. Maybe we should remove this day?

WEEKDAY VS. WEEKEND VOLUMES





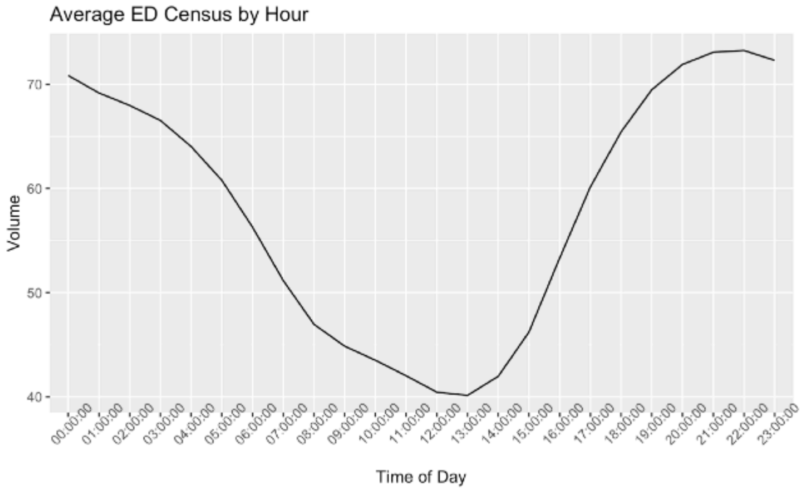
```
   "Weekday mean:"
   222.0426
   "Weekday standard deviation:"
   24.28034
   "Weekend mean:"
   213.1827
   "Weekend standard deviation:"
   19.06034

        Two Sample t-test

data:  n by day
t = 3.3297, df = 360, p-value = 0.0009593
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.627101 14.092786
sample estimates:
mean in group Weekdays mean in group Weekends
              222.0426               213.1827
```

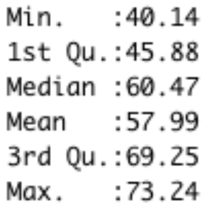
We can see from the Boxplot above that **there are about 9 more encounters on average on weekdays comparing to weekends**. To verify the significance of the difference we spotted, we perform a tow-sample t-test on the null hypothesis that average volume of encounters on weekdays is the same as weekends against the alternative hypothesis that they are different.

$$H_0 : avg_{weekday} = avg_{weekend} \text{ vs. } H_A : avg_{weekday} \neq avg_{weekend}$$

From the t-test result above, we can see that the p-value is p = 0.0009593, which is less than the significance level of 0.05. In conclusion, we strongly reject the null hypothesis. **The average volume on weekdays are significantly different form the average volume on weekends.**

CENSUS BY HOUR


Average ED Census by Hour



```
Min.   :40.14
1st Qu.:45.88
Median :60.47
Mean   :57.99
3rd Qu.:69.25
Max.   :73.24
```

As we can see, the mean census at any time of day is 57.99. On average, **census is the highest at around 21:00 to 22:00 each day and lowest at around 11:00 to 12:00 each day**.

SUMMARY

We can conclude that:

1. There are mostly 3 type of data quality problems: duplicate records, missing values and unrealistic times.
2. Both LOS and PIA_time are very right-skewed, with mean of 6.4 hours and 103 minutes respectively.
3. The most common presenting complaint is abdominal pain, followed by sore throat and loss of hearing.
4. The most common CTAS is urgent, followed by emergency and semi-urgent.
5. On average there are 221 encounters per day.
6. There are about 9 more encounter per day on weekdays comparing to weekends.
7. Census is hight from late afternoon til midnight (17:00 to 00:00).