|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Missing Data** | **Outcome** | | **P-value** |
| Alive | Die |
| Number of Patients | | 0 | 1465 | 278 |  |
| Urine output (mean) | | 63 | 1647.32 | 1147.42 | 1.46e-11 |
| Vasopressin usage during ICU stay (number of patients) | 0 not use(total=1574) | 0 | 1393 | 181 | 2.16e-21(和下面那一格一起算) |
| 1 use | 0 | 72 | 97 |  |
| Gender (number of patients) | Male | 0 | 744 | 132 |  |
| Female | 0 | 721 | 146 |  |
| hemoglobin\_min (mean+std) | | 1534 | 9.81+2.29 | 9.65+2.51 | 0.71 |
| hemoglobin\_max (mean+std) | | 1534 | 11.03+2.11 | 10.51+2.41 | 0.18 |
| resp\_rate\_mean (mean+std) | | 5 | 20.71+4.12 | 22.24+4.84 | 1.29e-6 |
| glucose\_max (mean+std) | | 1450 | 168.15+71.02 | 209.81+98.91 | 0.0032 |
| heart\_rate\_max (mean+std) | | 4 | 109.59+20.91 | 115.76+23.01 | 1.01e-5 |
| glucose\_avg (mean+std) | | 1450 | 148.91+54.55 | 186.34+86.32 | 0.0022 |
| Age (mean+std) | | 0 | 61.16+16.99 | 67.22+14.73 | 2.02e-9 |

­

表格中幾乎全部的數值全部都四捨五入到小數點第二位，沒有四捨五入到小數點第二位的數字是因為數字太小。

關於urine output，我發現在`physionet-data.mimiciv\_derived.urine\_output`中，它的charttime會記錄到比病人進icu的時間更早的資料，所以我直接在`physionet-data.mimiciv\_derived.urine\_output`中取時間從病人進icu前6小時到進入後24小時的總尿量。

關於p-value，我是在處理missing data和outlier前算的。因為number of patients沒有sample mean，所以沒有p-value，我就沒有算gender的p-value了。我的算法是先用stats.levene()，確認died和alive這兩群中的variance相不相近，再用stats.ttest\_ind()算p-value，看看這兩群的mean相不相近。如果variance相近，則在stats.ttest\_ind()中的equal\_var是True；如果variance不相近，則在stats.ttest\_ind()中的equal\_var是False。

Figures of features: .csv中的feature是gender加上以下圖片中的feature。

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

其中當vaso\_itemid=222315時，代表那位病人有用Vasopressin。

Data Preprocessing Description**、**Preprocessing Impact:

如果有Vasopressin usage，把vaso設為1；如果沒有Vasopressin usage，把vaso設為0。

經過把outlier(大於Q3+1.5\*IQR或 小於Q1-1.5\*IQR)設為mean，把missing data設為median的preprocessing處理之後，sample standard deviation有下降許多。例如：hemoglobin\_min的sample standard deviation從2.3左右下降成0.7左右，sum\_urine、hemoglobin\_max、glucose\_max也有下降。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Analysis:

一張含有 螢幕擷取畫面, 正方形, Rectangle, 文字 的圖片

自動產生的描述

我用corr() 算出每個feature之間的相關係數，再用heatmap去把corr()的結果顯現出來。

我先看hospital\_expire\_flag和其他哪一個features之間有比較大的關係，發現在ICU期間，有無Vasopressin usage和是否有死亡有較大的關係，當hospital\_expire\_flag變大時，vaso也比較會變大，也就是說病人的存活和有使用Vasopressin有正向的關係。

hospital\_expire\_flag與其它feature的相關係數都很接近0，所以比較沒有關係。

從這張圖，我發現到hemoglobin\_min和hemoglobin\_max之間的相關係數(0.71)很大，也就是說當hemoglobin\_max增加時，hemoglobin\_min也非常有可能增加。我也發現到glucose\_avg和glucose\_max之間的相關係數是0.55，也就是說當glucose\_avg增加時，glucose\_max有可能增加。

Note:

hw2的code是在hw2\_preprocessing\_data.ipynb和hw2.txt中。

此外，additional extract features寫在hw2\_optional.txt中。

關於additional extract features：

根據要求，我找出了Self-reported race, insurance, Congestive heart failure, Ventilation status, Lab values (max, min, avg): hemoglobin, glucose, lactate, 24 hr urine output。

除了24 hr urine output，其他跟時間有關的feature，我只取進ICU前6小時到進ICU後24小時之間的資料。

我找到了跟Congestive heart failure有關的17個icd\_code。如果病人有Congestive heart failure，icd\_code\_congestive\_heart\_failure, icd\_version\_congestive\_heart\_failure這2個column會呈現icd code, icd version，否則，就會呈現null。

一張含有 文字, 數字, 文件, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述