1.

a: 根據最後Model要output的**資料類型**去做決定—選擇應用classification model or regression model。而**資料類型**基本上分成兩類—quantitative or qualitative。Quantitative會以**數值**的型式呈現，如:BMI值的測量結果；Qualitative則以**分類**的型式呈現，並沒有數值的概念，如:動物照片的種族分類結果。Classification model的output結果屬於Qualitative的資料類型；而Regression model的output結果屬於Quantitative的資料類型。

根據題目所述，Model output資料屬性是Monthly rent，明顯屬於需要以**數值化的方式**表達的資料類型—Quantitative。故使用**Regression Model**會相較Classification Model更合適。

b: 以Unsupervised-learning的角度下，其主要解決的問題是Clustering—

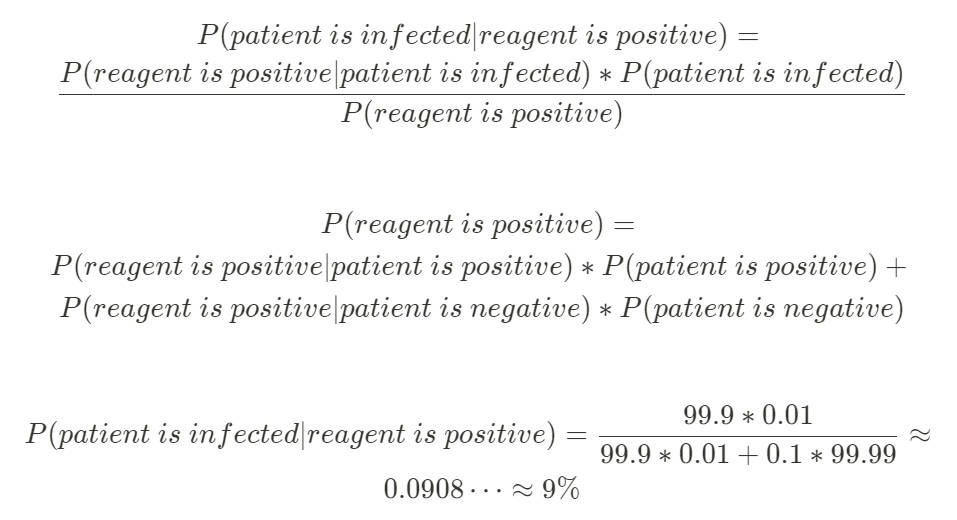
將**相似特性的資料聚集**，且**不相似的資料區隔/分開**。其類似於Supervised-learning的classification problem，同樣是要做區隔不同資料及聚集相似資料的動作。依我目前的知識推論，Unsupervised-learning的資料型式是以Qualitative的型式來呈現，若要表達數值化的區分，則做**數值範圍的區隔**，如BMI是23以上是肥胖症狀。雖然不能表達精確的數值化，但能表達其數值大約落在哪個區間。故我認為Unsupervised-learning針對此問題並無法有效的解決。

2.

HIV test 只有各別0.1%的機率會發生false positive and false negative。在特定群族之中，HIV的陽性盛行率(prevalence of HIV positive rate)是0.01%。

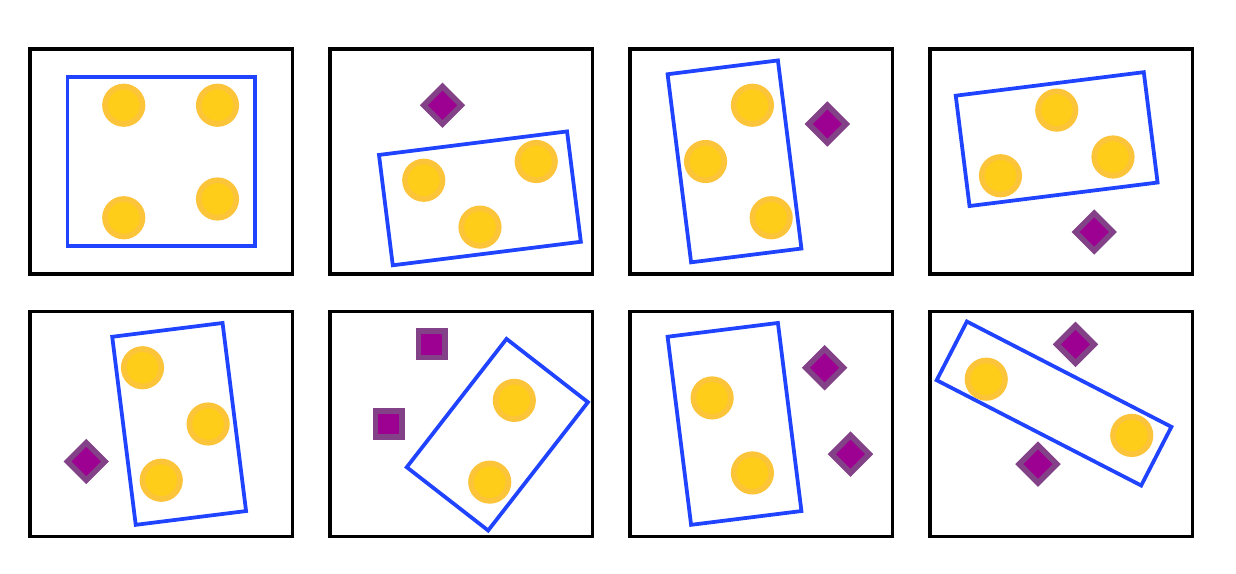
False positive 代表 P(reagent is positive | patient isn’t infected)—患者無染病，但卻誤判成positive (陽性)的情況。(題目2所述依我判別，False positive的寓意這樣合理)

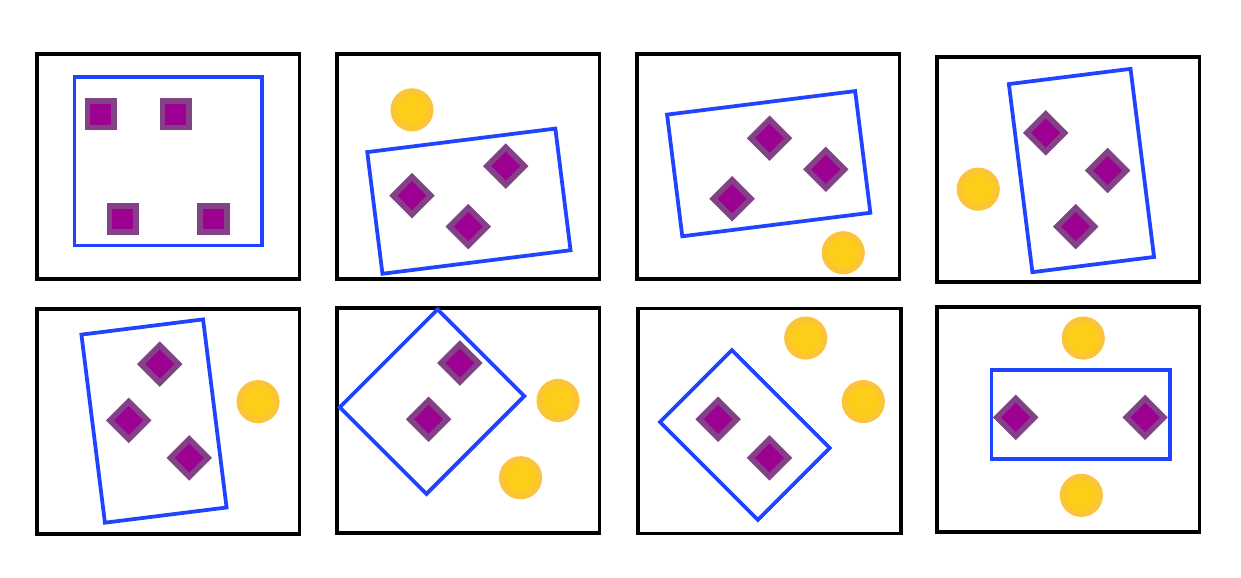
故題目中說明”若一個人屬於特定族群，且發現在HIV test中得到的結果為positive，那這人真正得到感染的機率為何?



3. Prove a rectangle decision function has VC dimension of 4 by enumerating all possible distributions.

The scenario of VC-dim = 4, dichotomy = 24 = 16:





4. 背面有附上程式碼，使用Jupyter Notebook Tool

5. 背面有附上程式碼，使用Jupyter Notebook Tool