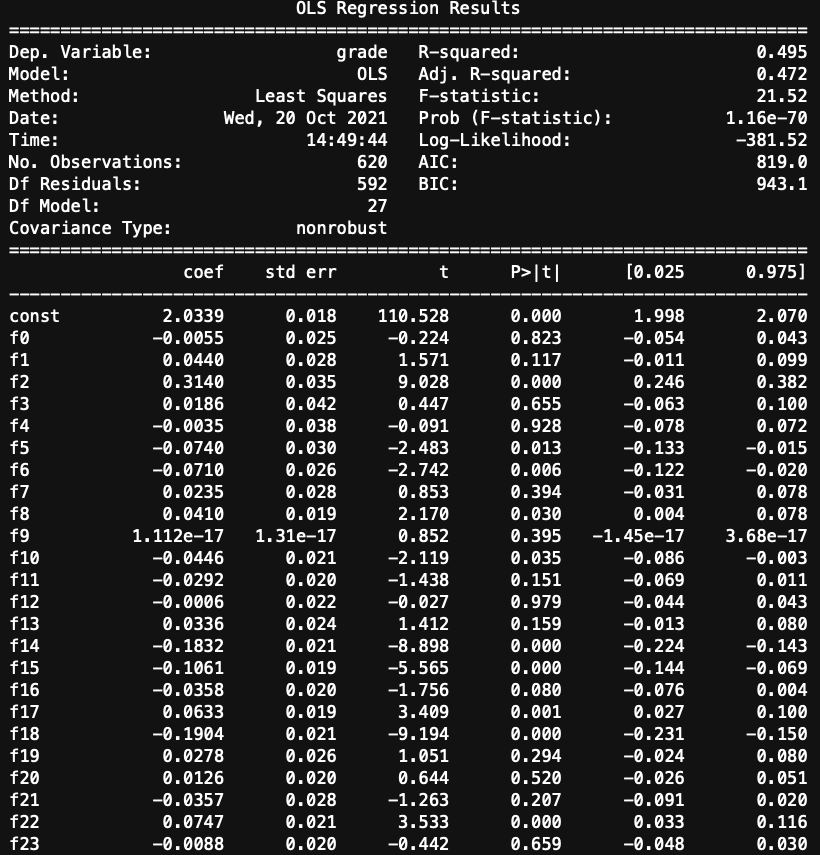
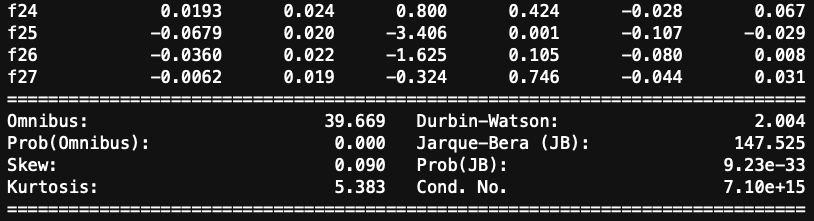
**Manufacturing Data Science製造數據科學**

**Assignment 1**

**姓名：張煜柔**

**學號：r08725059**

1. (a)



1. (b)

R-square = 0.495而Adjusted R-square = 0.472，與R-square並未相差太大，表示以目前的變數量而言維度詛咒並不明顯。然而因實務上並非所有案例其迴歸結果的R-square都很高，因此直接以目前的R-square斷言值太低、fit不好。客觀地說明則為由 R-square = 0.495可得此模型解釋49.5% grade的變異量。而若想增進此模型的R-square，可將自變數的交互作用考量進模型中，則有可能增加模型對grade變異量的解釋力。

1. (c)

f18 6.355895e-19

f2 2.430368e-18

f14 6.896905e-18

f15 3.971326e-08

f22 4.429125e-04

f17 6.967770e-04

f25 7.041639e-04

f6 6.295708e-03

1. (d)

參考助教Regression & ANOVA.ipynb內的教學[[1]](#footnote-1)，且由於資料筆數 < 5000，使用shapiro.test 函式來檢驗殘差的常態性。此檢定其虛無假設 為此資料來自於常態分佈，若p-value < 0.05 則可說在95%的信心水準下拒絕 。

其檢定結果為：test statistic = 0.9345124363899231、p-value = 7.536043131996446e-16，由於 p-value < 0.05，表示在95%的信心水準下拒絕殘差來自於常態分佈的虛無假設。

參考助教的教學，使用Durbin-Watson Test來檢驗殘差的獨立性。

根據statsmodels.stats.stattools.durbin\_watson的documentation[[2]](#footnote-2)，當test statistic = 2時，residuals間沒有serial correlation；當test statistic越接近0時，有更多證據證明residuals間有positive serial correlation；當test statistic越接近4時，有更多證據證明residuals間有negative serial correlation。

其檢定結果為：2.0040534804722108，以準則來看接近2，residuals間沒有serial correlation。

參考助教的教學，使用Breusch–Pagan test來檢驗殘差的變異數同質性。

根據statsmodels.stats.diagnostic.het\_breuschpagan的documentation[[3]](#footnote-3)，：the residual variance does not depend on the variables in x in the form。

其檢定結果為：

[('Lagrange multiplier statistic', 140.62034965182627),

('p-value', 5.87184293010542e-17),

('f-value', 6.43171099963128),

('f p-value', 5.3118156753530166e-20)]

由於p-value < 0.05，拒絕，這些殘差不具有同質性。

1. (a)

將根據變數的資料性質分為數值與非數值，分別進行敘述統計。

下表為數值欄位的敘述性統計：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | age | fnlwgt | education-num | capital-gain | capital-loss | hours-per-week |
| count | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 |
| mean | 38.581647 | 1.897784e+05 | 10.080679 | 1077.648844 | 87.303830 | 40.437456 |
| std | 13.640433 | 1.055500e+05 | 2.572720 | 7385.292085 | 402.960219 | 12.347429 |
| min | 17 | 1.228500e+04 | 1 | 0 | 0 | 1 |
| 25% | 28 | 1.178270e+05 | 9 | 0 | 0 | 40 |
| 50% | 37 | 1.783560e+05 | 10 | 0 | 0 | 40 |
| 75% | 48 | 2.370510e+05 | 12 | 0 | 0 | 45 |
| max | 90 | 1.484705e+06 | 16 | 99999 | 4356 | 99 |

下表為非數值欄位的敘述統計：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | workclass | education | marital-status | occupation | relationship | race | sex | native-country | income |
| count | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 | 32561 |
| unique | 9 | 16 | 7 | 15 | 6 | 5 | 2 | 42 | 2 |
| The most common value | Private | HS-grad | Married-civ-spouse | Prof-specialty | Husband | White | Male | United-States | <=50K |
| The most common value’s frequency | 22696 | 10501 | 14976 | 4140 | 13193 | 27816 | 21790 | 29170 | 24720 |

2. (3)

首先處理missing value的問題：

* 1. 檢視是否有空值。

檢查結果為所有自變數與應變數欄位皆沒有空值。

* 1. 檢視是否有「？」值。

根據Census Income Data Set說明文件[[4]](#footnote-4)提到：

| Conversion of original data as follows:

| 1. Discretized agrossincome into two ranges with threshold 50,000.

| 2. Convert U.S. to US to avoid periods.

| 3. Convert Unknown to "?"

| 4. Run MLC++ GenCVFiles to generate data,test.

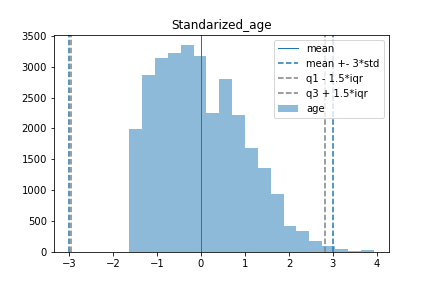
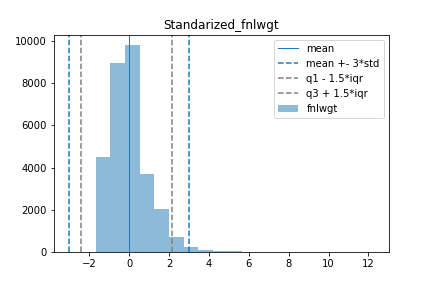
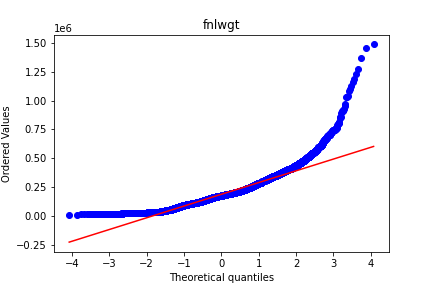
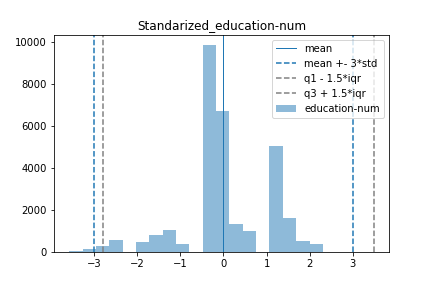
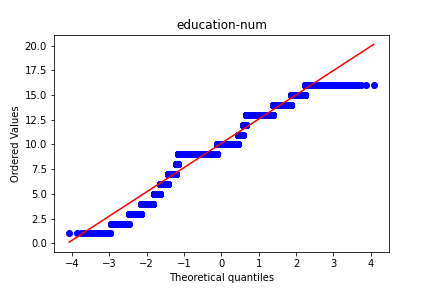
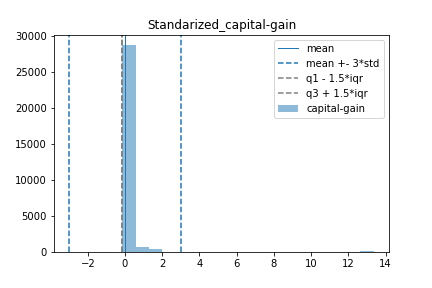
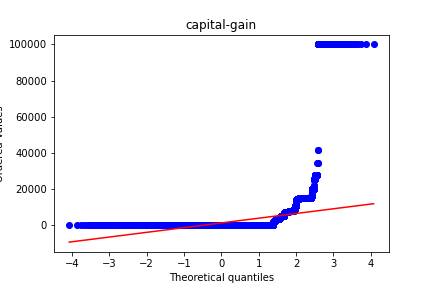
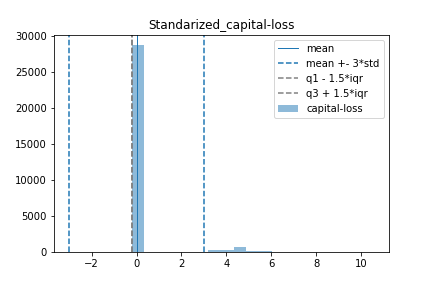
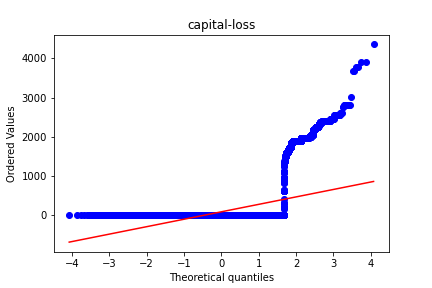
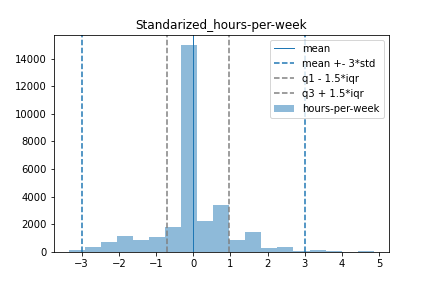
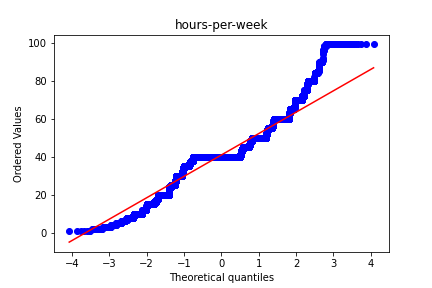
可知此data set的空值形式包含"?"，因此針對"?"值做missing value的處理。其中有空值的欄位其空值個數如下表：

|  |  |
| --- | --- |
| 自變數名稱 | 空值個數 |
| workclass | 1836 |
| occupation | 1843 |
| native-country | 583 |

接著將有包含空值的資料（列）移除，移除前的資料筆數為32561，移除後的資料筆數為30162。

接著針對numeric 自變數進行Outlier detection：

1. 首先繪製直方圖與QQ-plot檢視資料分布情況，並在**標準化後**的直方圖中標示出兩種判定outlier的方法：平均值加減三個表準差的範圍，與第一四分位數、第三四分位數加減1.5倍四分位距的範圍。可觀察到hours-per-week、education-num與age較接近常態分佈，其餘像是capital loss、capital gain則資料呈現較集中的分佈。

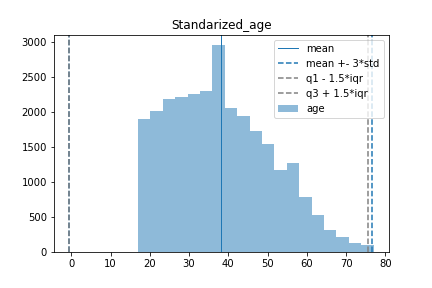
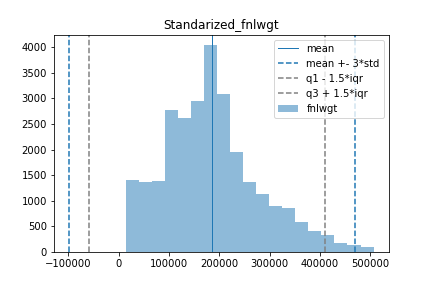
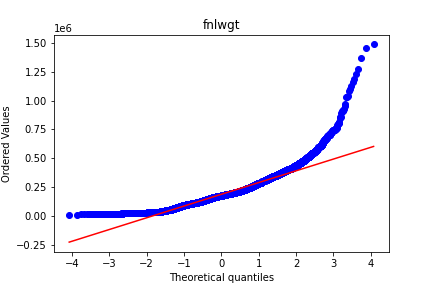
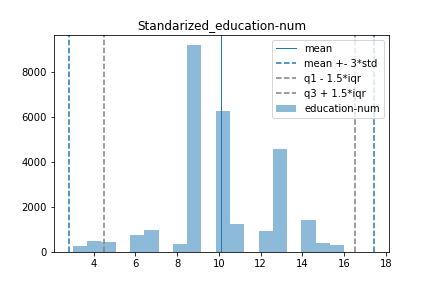
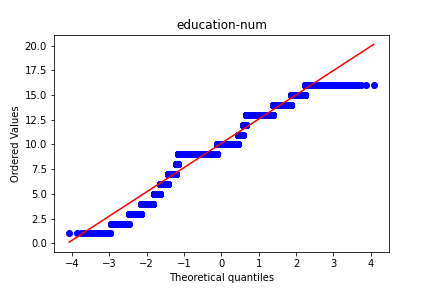
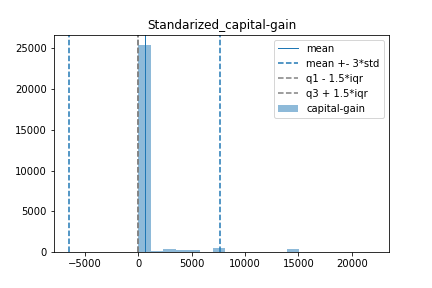
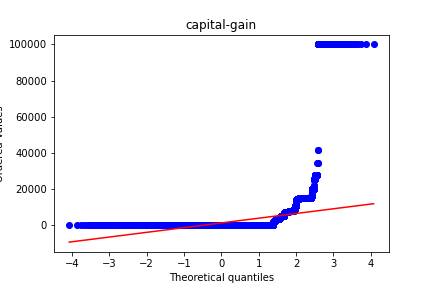
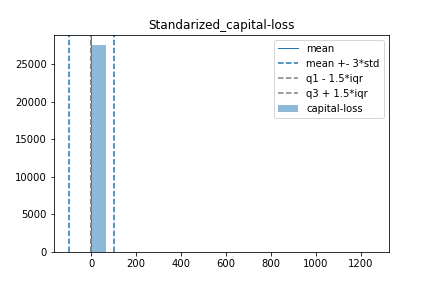
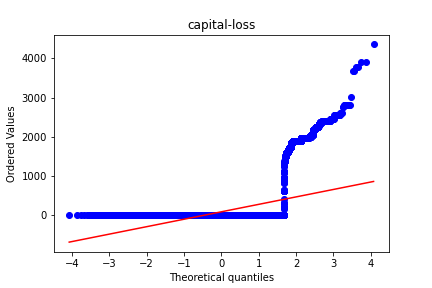
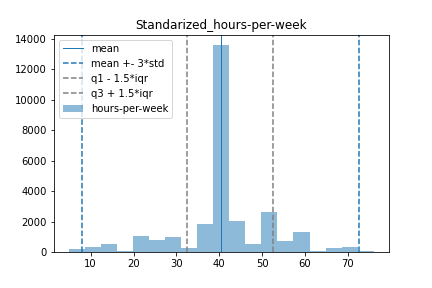
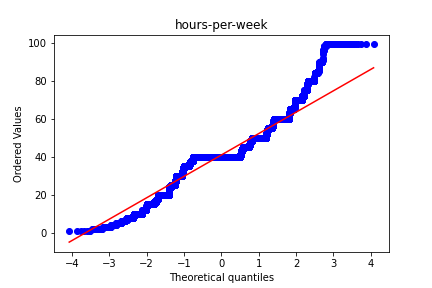


1. 檢視兩種判定方法下，被判定為outlier的資料數量：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | hours-per-week | capital-loss | capital-gain | education-num | fnlwgt | age |
| 平均值+-  3標準差 | 402 | 1381 | 198 | 196 | 322 | 120 |
| 四分位數+- 3四分位距 | 7953 | 1427 | 2538 | 484 | 903 | 169 |

相比之下，選擇平均值加減三標準差的outlier判斷方式，其保留較多資料筆數

1. 移除包含任一自變數outlier的資料，移除前的資料筆數為30162筆，移除後的資料筆數為27621筆。
2. 檢視移除outlier後的未標準化資料分佈。從上一小題的直方圖與QQ-plot中可觀察到部分自變數不服從常態分配，因此移除outlier後還是具有outlier，但可觀察到已移除較離群的值。



1. (3)

將變數進行get dummies處理，也就是若此變數有N個類別，則建立N-1個binary變數來呈現各資料在此變數的類別為何。自變數即執行pandas.get\_dummies(drop\_first=True) function，而應變數也同樣轉換成binary變數，將<=50K設為0，>50K設為1。

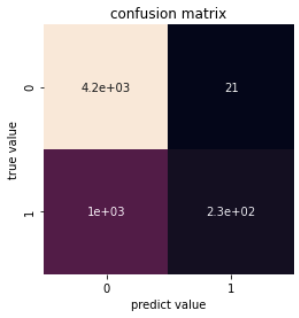
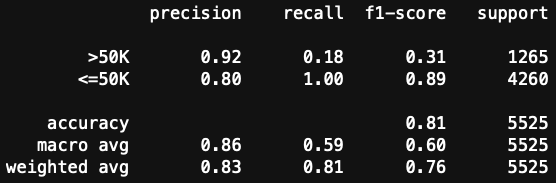
2. (4)

呼叫sklearn.model\_selection. train\_test\_split funtion，將資料以8:2比例隨機切成training data與testing data。

2. (5)

使用sklearn.linear\_model.LogisticRegression來進行預測。為增進模型的預測能力，使用training data進行parameters（probability threshold與c）tuning。訓練完成後預測testing data後的f1 score: 0.307、Accuracy score: 0.809。

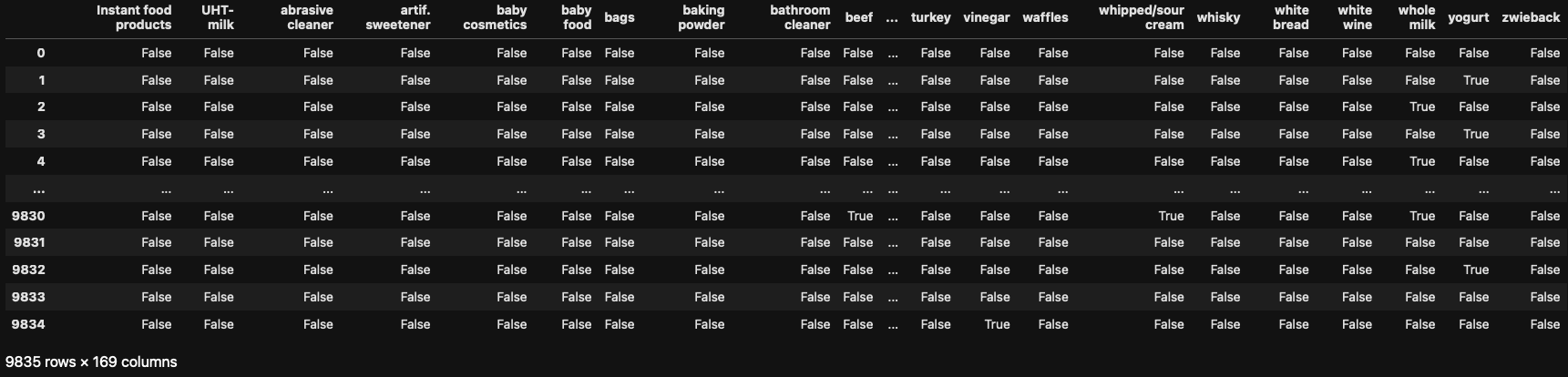
confusion matrix與classification report分別為：



綜合不同衡量指標可發現，此模型對於預測income >50K的資料有較好的預測效果。

1. (1)

使用mlxtend.preprocessing. TransactionEncoder將原始資料轉換為one hot encoding，並且存進dataframe資料格式，以下圖程式截圖為例：

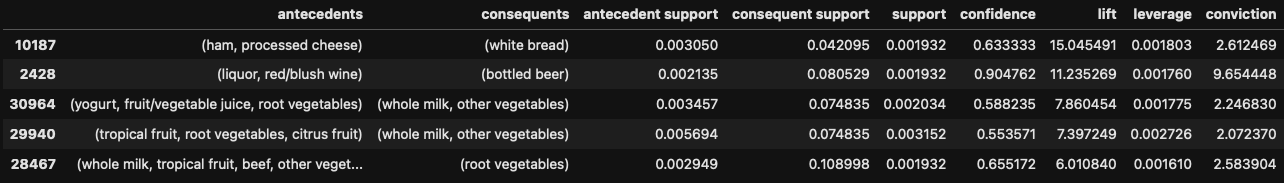


3. (2)

利用mlxtend.frequent\_patterns的apriori與association\_rules functions來建立關聯規則。並設定minimum support=0.001、minimum confidence=0.15。由於題目敘述希望找到top5的關聯規則，為避免找出的關聯規則都是普遍性產品，並且想找可靠性高、有效性也高的規則，先檢視前百分之十的規則其support、confidence與lift為何，如以下表格所示：

|  |  |
| --- | --- |
| Support | 0.002949 |
| Confidence | 0.586207 |
| Lift | 6.065211 |

因此，根據前百分之十的規則，先以support >= 0.0019、confidence >= 0.55以及lift >=6來進行規則的filter，filter出五個規則，並根據lift大小將其排序。其結果為：



3. (3)

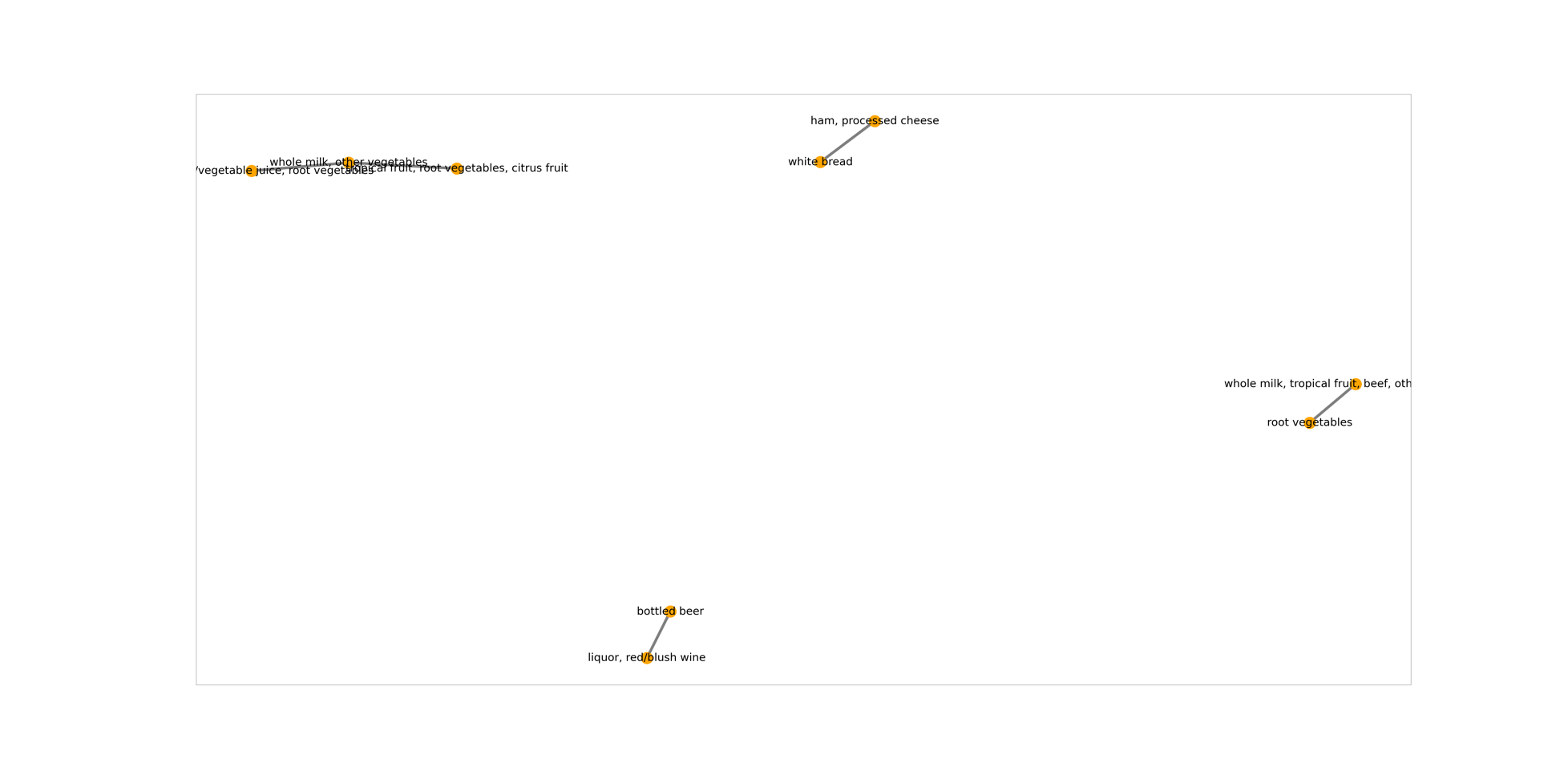
以top 1的規則為例，其support、confidence與lift對應到的機率分別為：

首先，ham、processed cheese與white bread同時出現的機率為，三者同時出現並不是非常大眾性的產品。而也有0.633左右的信心水準，相信購買ham、processed cheese下也會購買white bread。而Lift也有15，表示此規則的信賴度大於原本購買white bread的機率很多。因此可以相信顧客購買ham、processed cheese的同時，也會選購white bread。

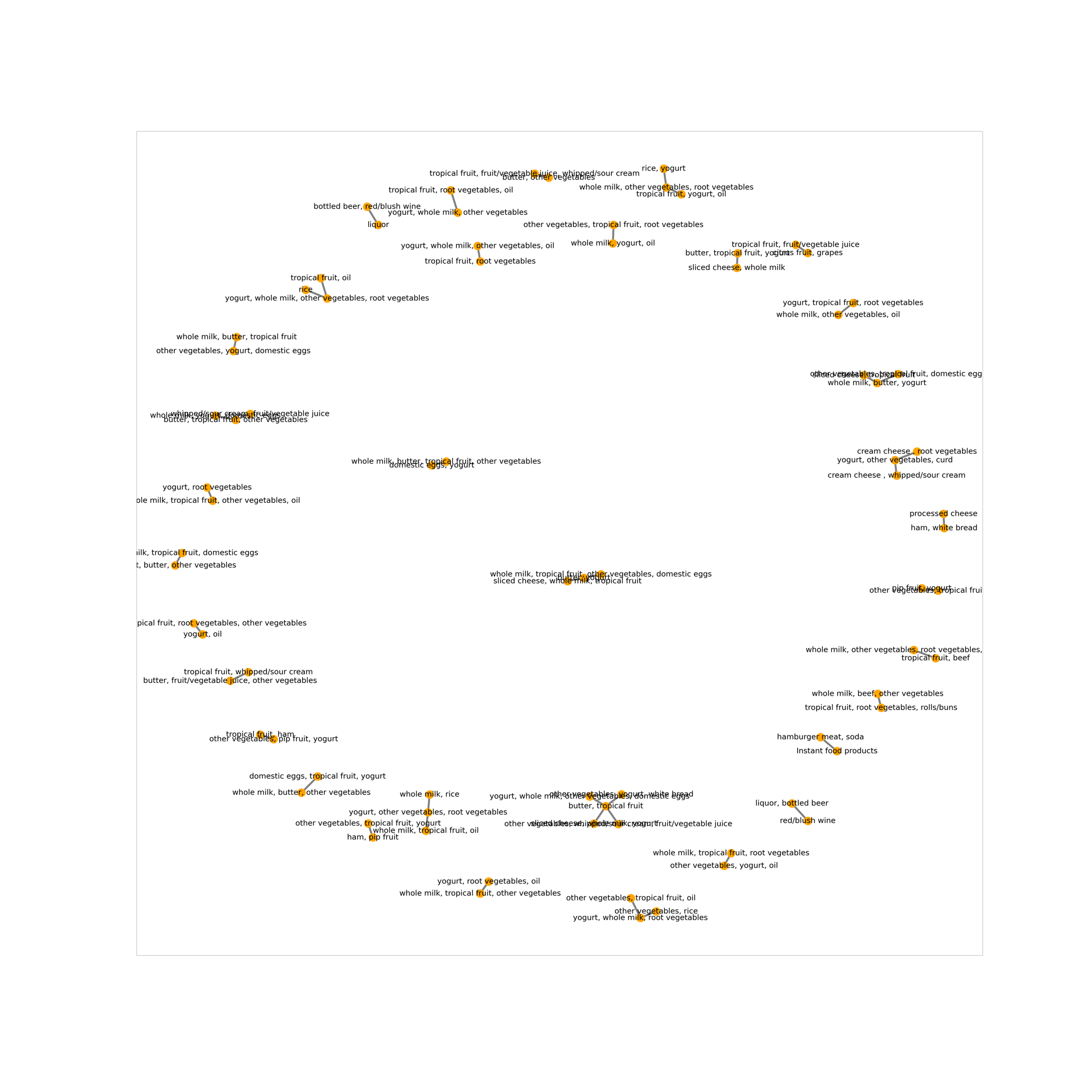
可猜測以下故事背景：通常火腿與起司，這類屬於較鹹的主食，不太會單獨食用，並且在國外通常會被製作成漢堡或三明治，比較簡單、方便，是國外常見的早午餐，因此可判斷此規則與現實情況相符。

3. (4)

以下為top 5規則的graph視覺化：



另外也會致根據lift進行排序的top 50規則的graph視覺化：



無論從top 5或top 50的規則圖都可發現有些產品的關聯性較高，在圖中則較相近，而較無關的產品則距離較遠，呈現產品間分群的現象。例如水果、根莖菜類、油、優格、牛奶在兩張圖的距離皆相近，對照到可能的食譜即為沙拉。

1. https://github.com/PO-LAB/Manufacturing-Data-Science/blob/master/MDS/04.%20Regression%20%26%20ANOVA/Regression%20%26%20ANOVA.ipynb [↑](#footnote-ref-1)
2. https://www.statsmodels.org/stable/generated/statsmodels.stats.stattools.durbin\_watson.html [↑](#footnote-ref-2)
3. https://www.statsmodels.org/stable/generated/statsmodels.stats.diagnostic.het\_breuschpagan.html [↑](#footnote-ref-3)
4. https://archive.ics.uci.edu/ml/datasets/Census+Income [↑](#footnote-ref-4)