



企業大數據分析師考試

AX01

情境手冊 Scenario Booklet

這是一個 2.5 小時的客觀型測試考試。

在前 50 分鐘之後，考生可以選擇兩次五分鐘的休息時間。休息時間必須要在回答完一個主要問題和所有相關的小問題之後。在每次休息期間 2.5 小時的考試時間將被暫停，休息結束之後再重新開始。考生必須管理好自己的時間以完成所有問題。

這本情境手冊包含了考試所依據的場景。所有問題都包含在問題手冊當中。

本情境手冊提供了一些問題的附加資訊。如果需要參考附加資訊，會在與之相關的問題中明確說明。每一個問題所提供的資訊，必須僅僅適用於該項問題當中。

考試最多有 80 分。每項題目各有 20 分。

及格分數為 65%（答對 52 項題目）。在每項問題中，都清楚地說明了該問題所涉及的課程大綱領域。

企業大數據分析師考試是 Open Book，考生可以在考試期間使用筆記型電腦、iPad 等閱讀官方的 EBDA 手冊。由於 EBDA 手冊可以在電腦內保存為 PDF 檔案，因此在考試開始前，所有設備都應處於飛行模式，以便阻止網路的使用。

候選人號碼：

此頁空白

場景中的公司和人員都是虛構的。

背景

保險公司 “全球綠色保險” (Global Green Insurance, GGI) 是世界領先的保險供應商之一，在五大洲擁有 2720 萬客戶。他們的核心重點是提供保險產品，以促進一個更可持續和更綠色的世界。例如，他們的領先產品之一是電動保險 (Electric Insurance)，這是一種專門針對電動汽車的保險產品，被認為是業內最具競爭力的汽車保險產品之一。

在過去的一年裡，GGI 在擴大客戶群方面遇到了重大挑戰。雖然在廣告方面花費了數百萬，但幾乎沒有增長。目前還不太清楚是什麼原因導致了這個問題，幾乎整個執行團隊都在質疑廣告宣傳費用是否得到了有效的使用。此外，全球綠色保險在過去兩年中面臨著一些新的競爭者，這些競爭者已經慢慢獲得了信譽並開始成為 GGI 的威脅。儘管這些新的競爭者的資源遠不及 GGI，但他們在瞄準新客戶方面卻更加有效。

高階執行團隊

為了實現進一步增長，董事會最近任命了一位在技術和數據分析方面具有豐富經驗的新首席執行官。CEO 的願景是將 GGI 轉變為數據驅動型組織。已經建立了一個由 20 名數據分析師組成團隊的大數據卓越中心 (Big Data Centre of Excellence)，並任命了首席數據官。

數據

作為擁有 2720 萬客戶的全球組織，資訊部門的業務範圍很廣。有關客戶、產品和索賠的數據被廣泛地被收集，但是 95% 的數據從未在任何類型的分析中被使用過。

GGI 與供應商建立了戰略合作夥伴關係，提供外部數據的使用。該公司購買有關社群媒體使用、潛在新客戶和廣告活動的數據。

GG 的 “核心” 系統之一是集中型的 SQL 資料庫，其中包含公司曾經出售給客戶的所有保單。數據儲存在大型資訊供應商的專屬型的資料庫解決方案中。該資料庫包含許多有關客戶關鍵特徵的數據，例如年齡、性別、位置和針對特定保單的索賠數量等。新的首席數據官已經指出，該資料庫可以視為組織的 “金庫”。

場景說明結束

題目1, Part B 的附加資訊
Additional Information for Question 1, Part B

根據首席執行官的指示，首席行銷官(CMO) 已經開始了一個專案，建立一個模型，預測廣告支出的投資回報率(ROI)。為了進行任何分析，她首先需要確保數據被導入到新的數據分析工具中。

要求一名數據分析師整合來自以下來源的數據：

- 1. 第三方的保險銷售資訊：“InsuranceSales.csv”，其第一行包含有標題。
- 2. 由外部供應商提供的，關於廣告活動點擊率的資訊：“Click_Rate.txt”。這個檔案不包含任何標題。
- 3. 顯示人們在流行的社群媒體平台：“Friends4Ever”上對GGI產品的評價的資訊，可利用應用程式接口(API) 進行讀取。

```
<Root xmlns="http://www.friends4ever.com">
  <Customers>
    <Customer CustomerID="CUSTOMER-1">
      <CompanyName>Car Insurance Company</CompanyName>
      <ContactName>Howard Snyder</ContactName>
      <ContactTitle>Marketing Manager</ContactTitle>
      <Phone>(503) 555-7555</Phone>
    </Customer>
    <Customer CustomerID="CUSTOMER-2">
      <CompanyName>Global Insurance Corp</CompanyName>
      <ContactName>Yoshi Latimer</ContactName>
      <ContactTitle>Sales Representative</ContactTitle>
      <Phone>(503) 555-6874</Phone>
      <Fax>(503) 555-2376</Fax>
    </Customer>
  </Customers>
</Root>
```

Figure 1: Example of “Electric-Insurance” data that needs to be imported

圖 1：需要被導入的 “Electric-Insurance” 數據案例

- 4. 除了上面列出的來源之外，CMO 還收到了一個JSON 檔案的鏈接，該檔案包含有關由合作夥伴組織在網路上銷售的GGI “Electric-Insurance” 保單的資訊。
- 5. 從“core” 全球綠色保險SQL資料庫中的表格中提取的客戶數據。

題目1, Part C 與 Part D 的附加資訊
Additional Information for Question 1, Part C and Part D

“InsuranceSales.CSV” 數據已經導入GGI 的新數據分析工具中。為了確定所考慮的數據集，數據分析人員已進行了一些探索型數據分析(exploratory data analysis)。

數據分析師首先檢查了數據集的結構：

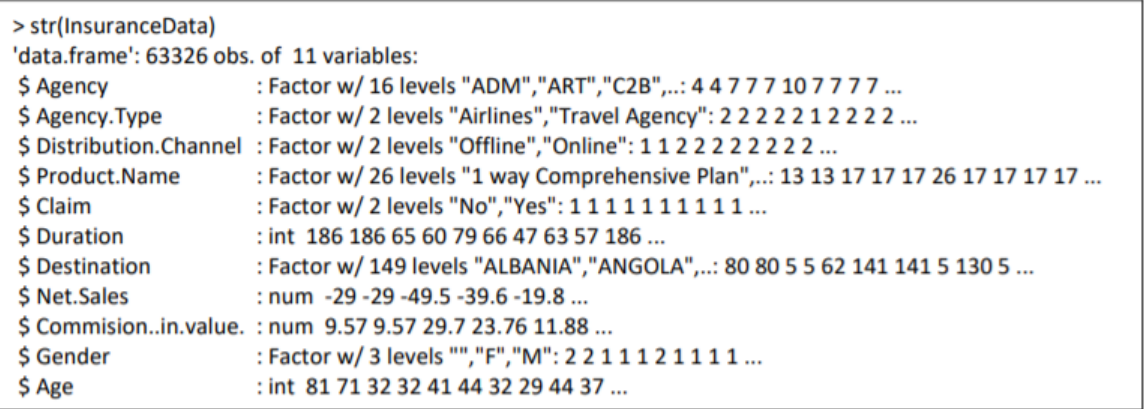


Figure 2: Exploratory Data Analysis of the Insurance Sales

圖2：Insurance Sales 的探索型數據分析

然後，數據分析師檢查摘要統計資訊：

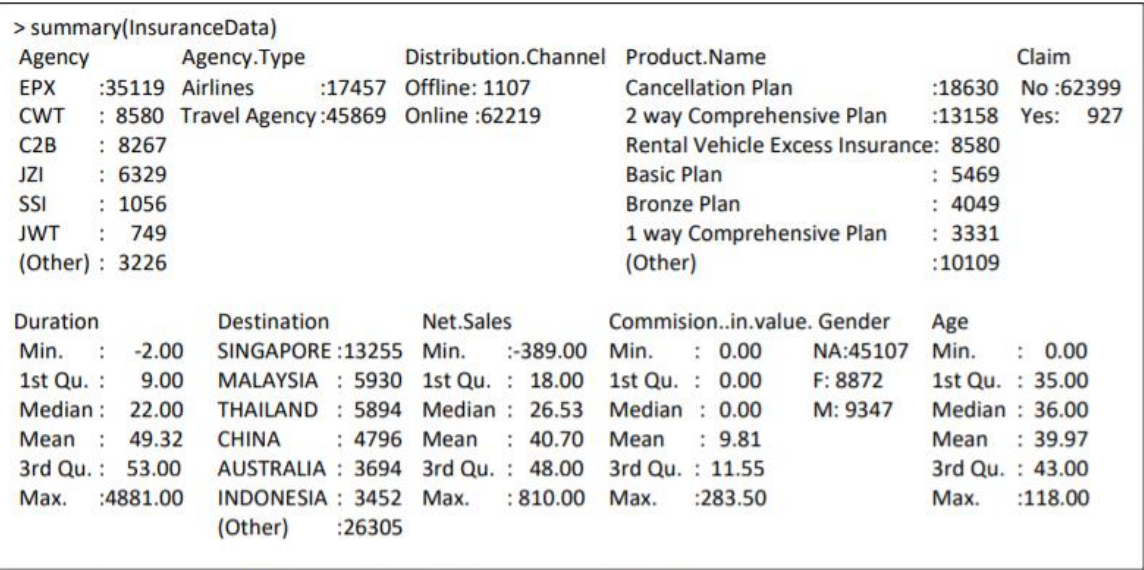


Figure 3: Exploratory Data Analysis of the Insurance Sales

圖3：Insurance Sales 的探索型數據分析

題目2, Part A 的附加資訊
Additional Information for Question 2, Part A

首席產品官想要知道，向男性和女性銷售保險產品是否有什麼不同。他還在考慮是否應該為 30 歲以上的女性專門設計一種新的保險產品。

來自 Insurance.Sales.CSV 檔案的保險銷售數據已經被導入並清理。數據分析師現在已經提取了以下探索型數據分析的圖形：

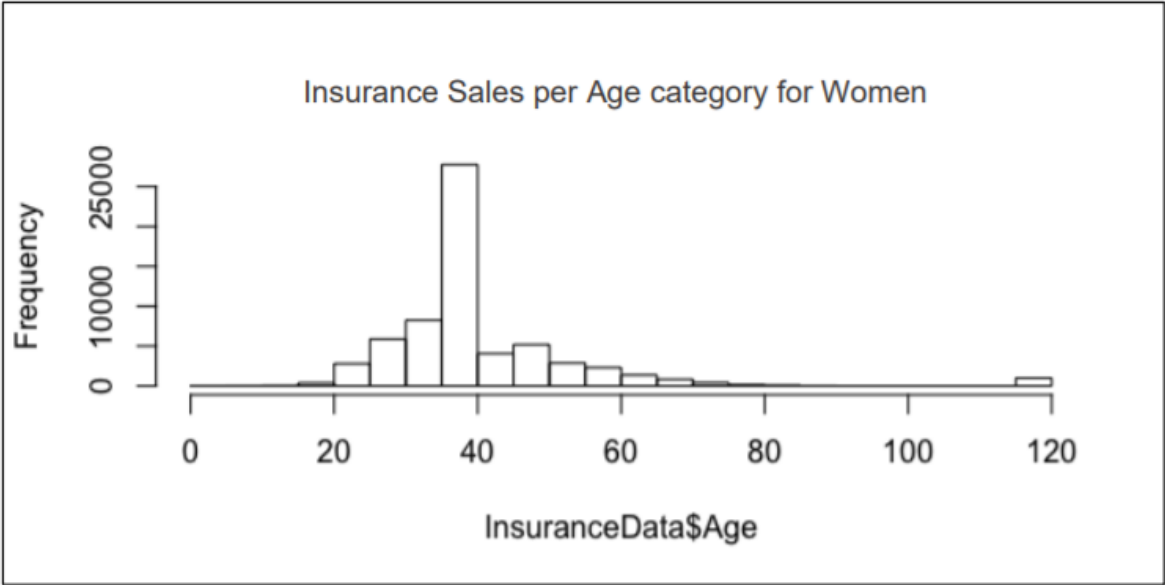


Figure 4: Insurance Sales per age category for Women

圖 4：按 Age 分類的女性保險銷售

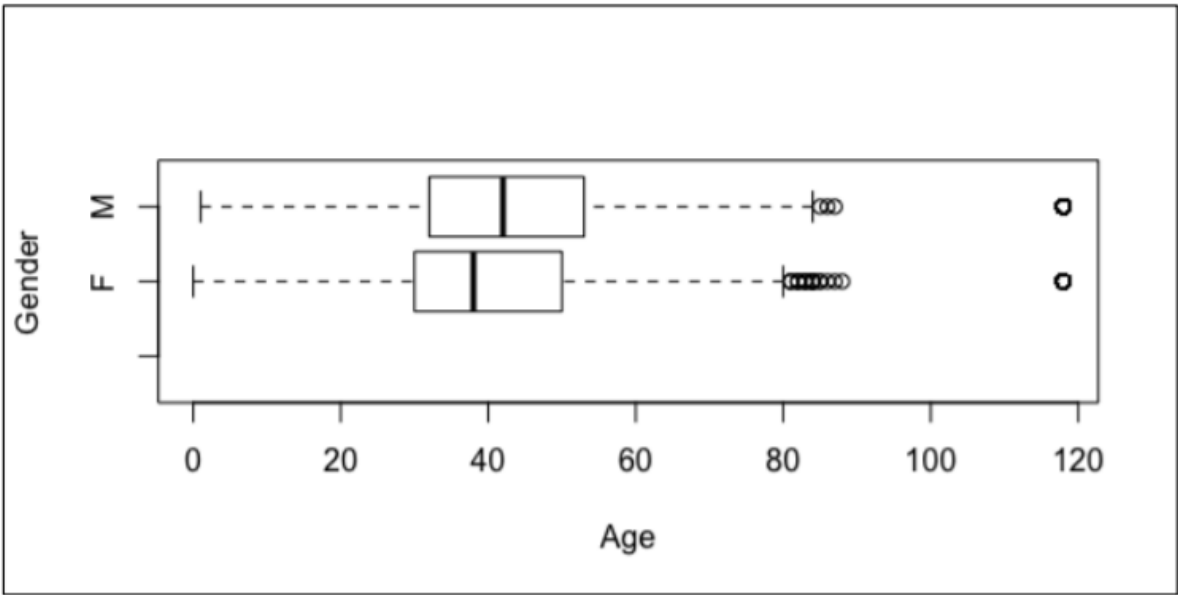


Figure 5: Insurance Sales per age category, split on the gender variable

圖 5：按 Age 分類的保險銷售，依據 Gender 變量來劃分

題目2, Part A 的附加資訊 – 繼續

Additional Information for Question 2, Part A - Continued

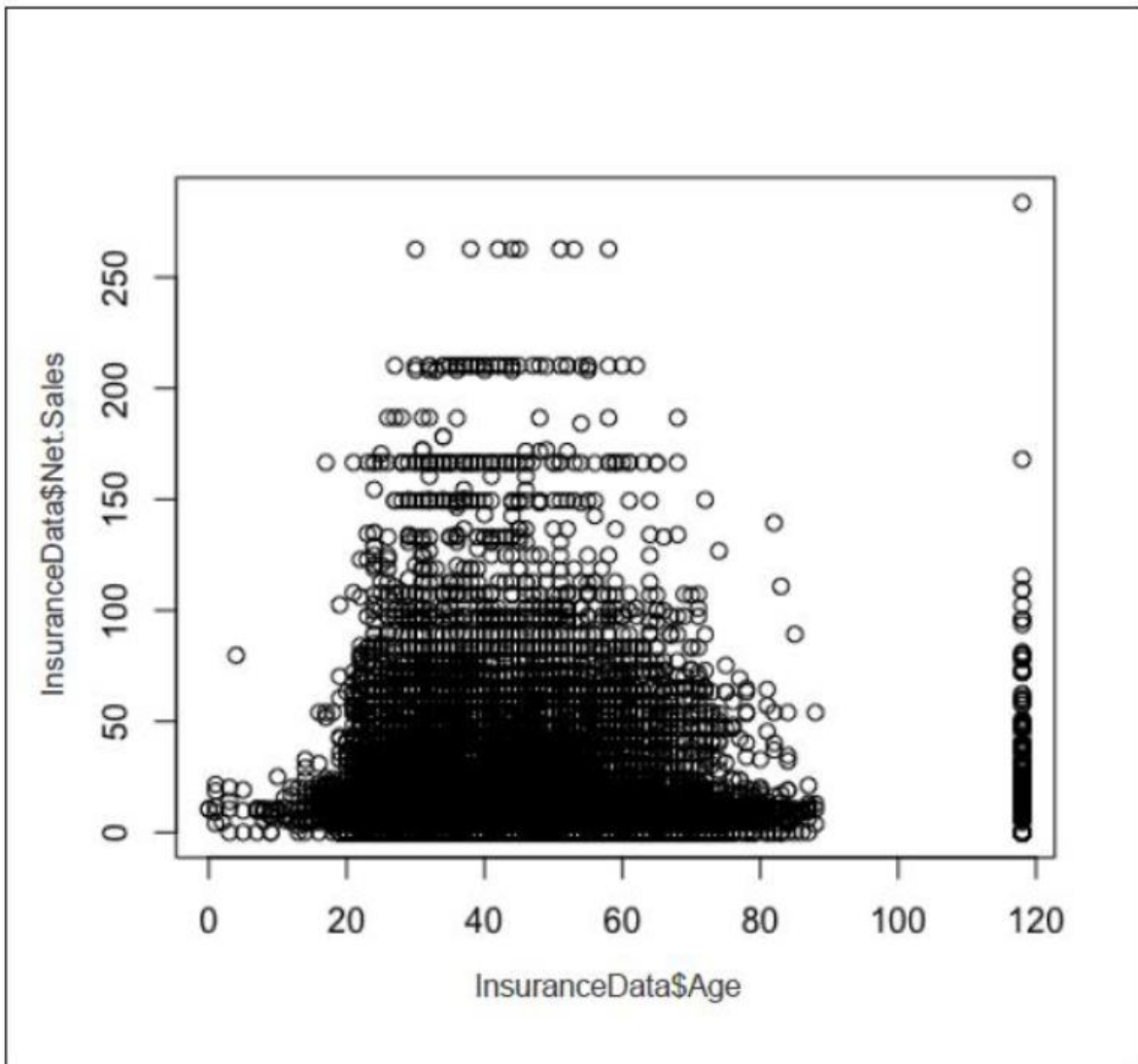


Figure 6: Insurance Sales Commission as a function of Age.

圖 6 : Insurance Sales 與 Age 的關係

題目2, Part B 的附加資訊
Additional Information for Question 2, Part B

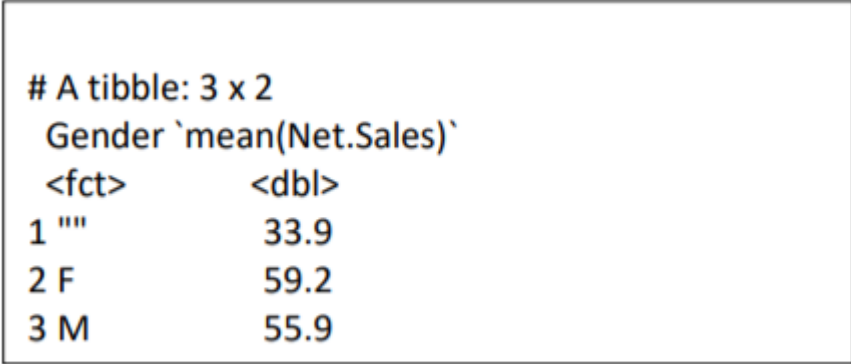
在完成探索性數據分析後，數據分析師懷疑女性比男性願意在保險上花更多的錢，因此設計一種專門針對女性的汽車保險產品服務可能是個好主意。為了確定這個懷疑是否正確，他提出了以下兩個假設。

- H0 --男女在保險上的平均花費是一樣的
- HA --女性在保險方面的平均支出高於男性

第一步，數據分析師的目標是找到觀察到的差異統計量(difference statistic)。為了計算這一統計量，他使用以下函數：

```
InsuranceData_Grouped <- group_by(InsuranceData, Gender)
summarize(InsuranceData_Grouped, mean(Net.Sales))
```

這產生了下表：



```
# A tibble: 3 x 2
  Gender `mean(Net.Sales)`
  <fct>      <dbl>
1 ""         33.9
2 F          59.2
3 M          55.9
```

Gender	mean(Net.Sales)
""	33.9
F	59.2
M	55.9

Figure 7: Mean sales, split on the “Gender” variable

圖 7: Mean sales, 按 “Gender”變量劃分

然後，數據分析師使用R中的“infer”套件，對保險銷售數據集的100個不同隨機樣本測試零假設。所得的100 個數據點的第95 個百分點低於觀察到的差異統計量。

題目2, Part C 的附加資訊

Additional Information for Question 2, Part C

數據分析師渴望理解他是否能找到GGI保險產品的“Net.Sales”和銷售保險產品的“Duration”（以月為單位）之間的關係。利用這些資訊，他的目的是在確定針對年輕人的保險產品是否比針對老年人口的產品更有利可圖。

他產生的輸出圖如下所示：

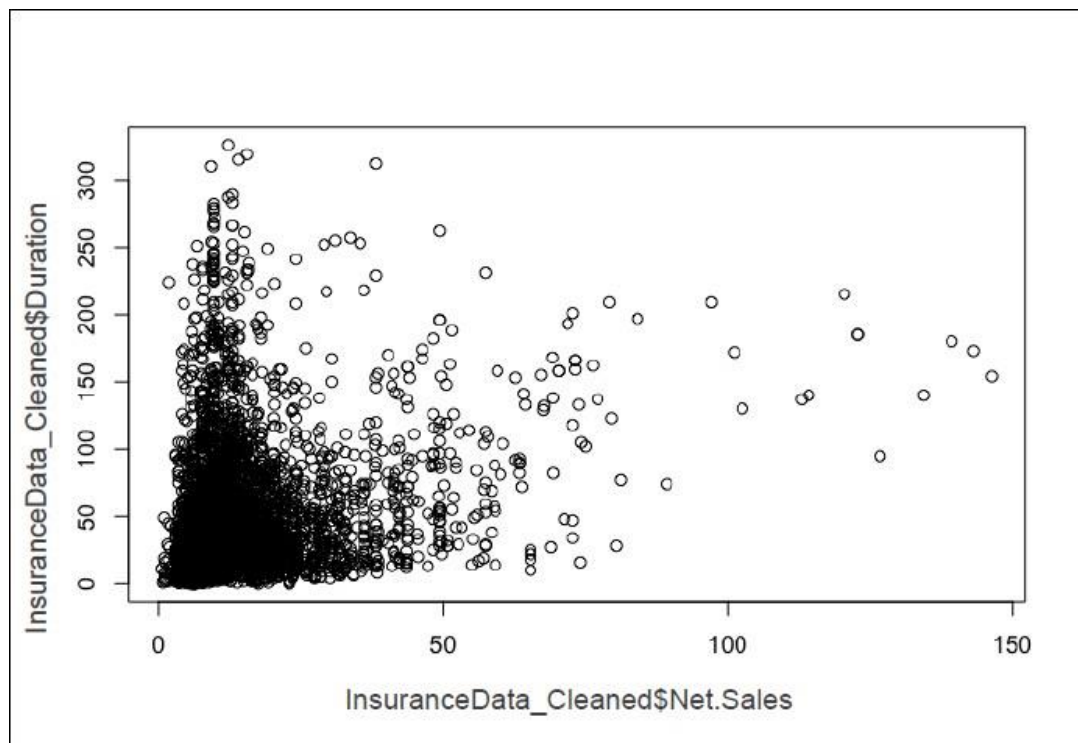


Figure 8: Insurance duration as a function of net sales value.

圖 8：Insurance duration 與 net sales 的關係

相應的 Pearson 關聯係數已計算如下：

```
> cor(x = InsuranceData_Cleaned$Net.Sales., y = InsuranceData_Cleaned$Duration)
[1] 0.3143053
```

題目3, Part A 的附加資訊
Additional Information for Question 3, Part A

除了研究“性別Gender”、“年齡Age”和“保險期限Insurance Duration”這些變量是否會影響客戶購買的保險產品類型，CMO還詢問是否有可能做一些更深入的數據分析形式，將客戶分成不同的群體。

作為後續工作，數據分析團隊已經確定，分類演算法可以用來解決這個問題。他們根據公司目前掌握的關於地理位置、人口統計學和購買行為(geographical location, demographics and purchasing behaviours)的數據，想出了一些不同的方法。

客戶已經根據他們的年消費額被分為3個消費等級。“高級Premium”、“標準Standard”和“經濟型Budget”。將為每種等級制定不同的行銷策略。

1) k-NN 分類(k-NN Classification)

數據分析團隊正在開發一個可以預測新客戶和潛在客戶的消費等級的模型。他們從客戶數據中選擇了五個屬性作為輸入數據，並將年度消費等級（“保險類別Insurance Class”）作為目標變量。有63,326個觀察值。

第一個k-NN模型使用了以下操作。

```
Insurance_Training <- InsuranceData[1:60000,]  
Insurance_Test <- InsuranceData[60001:63326,]
```

然後為測試數據產生年度支出類別（“保險類別Insurance Class”）的預測值，並與原始測試數據中的數值進行比較。

```
InsuranceData_Predicted <- knn(Insurance_Training, Insurance_Test,  
Insurance_Class)  
  
> [1] 0.45857
```

然後，該模型被多次運算，每次都增加用於計算歐幾里得距離的數據點的數量。

k=	Model accuracy
3	0.48972
5	0.58997
7	0.78223
9	0.57891

題目3, Part B 的附加資訊

Additional Information for Question 3, Part B

2) 單純貝葉斯(Naïve Bayes)

一位數據分析師產生了一個Naïve Bayes分類模型，以預測客戶是否會根據某些人口統計標準來購買GGI保險產品。

他已經建立了這個模型，現在正在審查概率表（詳見下文摘錄）。

===== Naïve Bayes =====	
Call: naive_bayes.formula(formula = Product.Name ~ Gender + Age, data = InsuranceData)	

A priori probabilities:	
1 way Comprehensive Plan	2 way Comprehensive Plan
0.0526008275	0.2077819537
24 Protect	Annual Gold Plan
0.0039004516	0.0030635126
Annual Silver Plan	Annual Travel Protect Gold
0.0224710230	0.0015791302
Annual Travel Protect Platinum	Annual Travel Protect Silver
0.0008369390	0.0013580520
Basic Plan	Bronze Plan
0.0863626315	0.0639389824
Cancellation Plan	Child Comprehensive Plan
0.2941919591	0.0001421217
Comprehensive Plan	Gold Plan
0.0057480340	0.0055585384
Individual Comprehensive Plan	Premier Plan
0.0011685564	0.0030635126
Rental Vehicle Excess Insurance	Silver Plan
0.1354893725	0.0355146385
Single Trip Travel Protect Gold	Single Trip Travel Protect Platinum
0.0032214256	0.0011527651
Single Trip Travel Protect Silver	Spouse or Parents Comprehensive Plan
0.0027318953	0.0002368695
Ticket Protector	Travel Cruise Protect
0.0166756151	0.0083220162
Travel Cruise Protect Family	Value Plan
0.0000157913	0.0428733853

Figure 9: Overview of the probability tables that will be used to make the classification predictions (extract only shown)

圖9：將用於進行分類預測的概率表的概述（僅顯示摘錄）

題目3, Part C 的附加資訊

Additional Information for Question 3, Part C

首席數據官對新的大數據卓越中心到目前為止所提出的分析印象深刻，並渴望了解更多。為了確保公司對其保險產品進行了充分的承保，數據分析團隊被要求創建一個模型來預測客戶是否會對其保險進行索賠，是或不是。

首先，一個數據分析師載入了數據，指定了最有可能產生預測影響的變量。下表顯示了在樣本數據集中有多少人進行了索賠。

>table(Claim_Data\$Claim)	
0	1
62399	927

在63,326個GGI客戶的樣本數據集中，大約1.5%的客戶已為其保險提出索賠。GGI在全球擁有2720萬客戶。能夠預測最有可能提出索賠的前1.5%或更多的客戶，並且能夠定位並可能減少該數量，這對GGI而言是巨大的機會。

然後，數據分析師使用相同的變量，構建了邏輯回歸模型：

```
Claim_Model <- glm(Claim ~ Product.Name + Duration + Gender + Age, data =  
Claim_Data, family = 'binomial')  
Summary(Claim_Model)
```

數據分析師將預測操作應用於此模型，以計算客戶提出索賠的概率（以百分比表示）。

```
Claim_Data$Predict <- predict(Claim_Model, type = "response")  
head(Claim_Data)
```

依據預測的概率得分對結果進行排序時，樣本數據集中前1.5%的客戶得分為0.2189或更高。數據分析師使用此得分0.2189作為將客戶分類為潛在“索賠人”的閾值。

```
Claim_Data$Predict_Claim <- ifelse(Claim_Data$Predict >= 0.2189, 1, 0)  
head(Claim_Data)
```

題目3, Part D 的附加資訊

Additional Information for Question 3, Part D

數據分析團隊根據現有客戶的情況和購買行為，建立了一個分類樹模型，以預測新客戶最有可能購買GGI的16種保險產品。分析模型的結果將被報告給高層執行團隊。

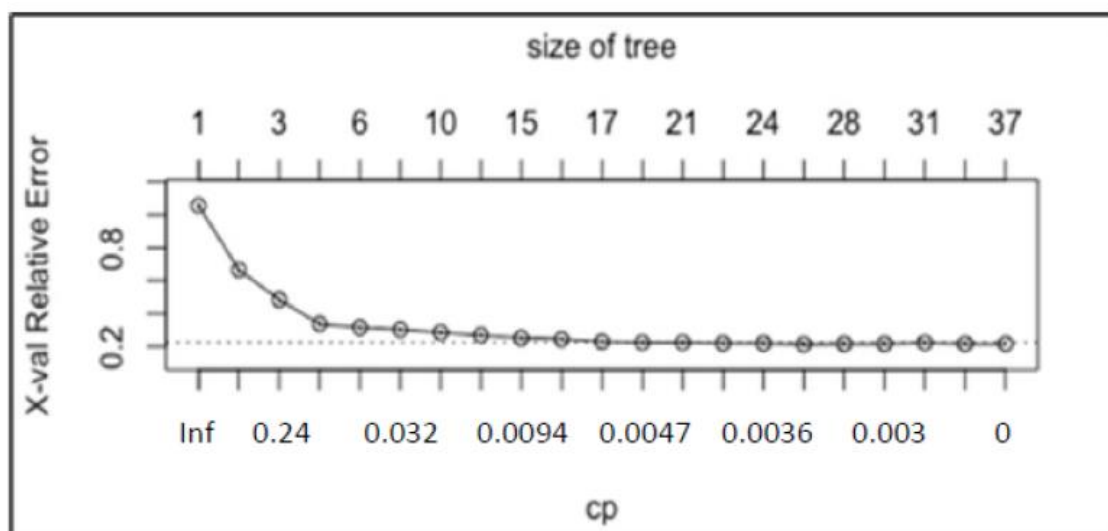
為了建立模型，相關資訊從GGI客戶的SQL數據庫中導入，作為一個有63,326個觀察值和11個變量的數據集。其中的一個子集被分離出來作為測試集使用。該模型的第一次執行是利用R語言的以下操作產生的。

```
Insurance_Data<- rpart(Product_Name ~ ., data = Customer_Train, method = "class",  
control = rpart.control(maxdepth = 10))
```

然後使用測試數據集來評估第一個模型的準確性：

```
Customer_Test$Predict <- predict(Customer_Model_Pruned, Customer_Test,  
type = "class")  
mean(Customer_Test$Predict == Customer_Test$Insurance_Product)  
[1] 0.829
```

然後構建第二個未修剪的模型，並使用plotcp() 函數分析樹的複雜度以生成以下圖形：



此頁空白



ENTERPRISE BIG DATA ANALYST

企業大數據分析師考試

AX01

問題手冊 Question Booklet

候選人號碼:

此頁空白

所涵蓋的課程大綱範圍

題目 1 - 提取與準備 (Ingestion and Preparation)

題目 2 - 探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)

題目 3 - 分類模型 (Classification Models)

題目 4 - 聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
提取與準備 (Ingestion and Preparation)	1	A	5

回答以下有關 GGI業務目標的問題。

第 1列是 GGI希望回答的問題列表。對於第 1列中的每項 問題，從第 2列中選擇需要找到相對應 業務目標類型，成為你的答案。每項 第 2列中的選擇可以使用一次、多次或完全不使用。

	第 1列	第 2列
1	GGI 的汽車保險銷售增長是否因為 (because of) 客戶群老化而受到限制?	A 描述型業務目標 (Descriptive) B 探索型業務目標 (Exploratory) C 推理型業務目標 (Inferential) D 預測型業務目標 (Predictive) E 因果型業務目標 (Causal) F 機理型業務目標 (Mechanistic)
2	購買 Electric-Insurance 的客戶通常會購買其他保險產品嗎?	
3	根據從以前的市場 活動中收到的客戶反饋樣本 (sample)，對於所有 GGI 客戶來說，什麼可能是最佳行銷方法?	
4	GGI 客戶的主要特徵 (key characteristics)是什麼?	
5	為什麼 (Why is) Electric-Insurance GGI 是主要 領先的保險產品?	

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
提取與準備(Ingestion and Preparation)	1	B	5

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關導入和讀取 廣告支出模型數據的問題

對於以下每種情況，請確定所採取的措施是否適當，然後選擇可以支持您所決定的選項

1	數據分析師編寫了以下程式碼，將 InsuranceSales.csv 檔案導入 R file <- read.csv(“InsuranceSales.csv”, header = FALSE) 利用這個程式碼，從這個檔案中導入數據是否正確？	<p>A 否，因為此數據集的第一行 (first row) 包含列標題(column headers)</p> <p>B 否，因為該公司的核心 GGI 資料庫為 SQL 格式</p> <p>C 是，因為在 R 中工作時會使用 read.csv() 命令</p> <p>D 是，因為 R 無法讀取數據，除非是 CSV 格式</p>
2	數據分析師編寫了以下程式碼，來下載 Click_Rate.txt 檔案。 file <- read.lines(“Click_Rate.txt”, skip = 0) 利用這個程式碼，從這個檔案中導入數據是否正確？	<p>A 否，因為默認(default)的情況下標頭(header)將被指定為 true</p> <p>B 否，因為 read.lines 不是將文字檔案導入 R 的有效命令</p> <p>C 是，因為外部檔案全部基於文字且易於閱讀</p> <p>D 是，因為應該先明確指定檔案名稱，然後說明沒有標頭(header)</p>
3	數據分析師計劃使用 download.file() 函數從 “Friends4Ever” 社群媒體平台下載數據，如圖 1 所示。 這種方法對這種檔案格式來說是否正確？	<p>A 否，因為 “Friends4Ever” 使用於數據沒有標記數據(no markup data)</p> <p>B 否，因為 “Friends4Ever” 是使用於 Excel 編碼的檔案</p> <p>C 是，因為可以使用 download.file() 方法從 Internet 下載所有類型的檔案</p> <p>D 是，因為可能需要使用 API 來讀取 “Friends4Ever” 數據</p>
4	鑑於數據分析師使用 R 程式語言工作，應用於合作夥伴出售保單的詳細資訊之 JSON 檔案，是否是一種可接受的數據儲存和交換的格式？	<p>A 否，因為組成 JSON 檔案的字元(characters)，分為標記(markup)和內容(content)</p> <p>B 否，因為 JSON 是一個開放源代碼的關聯資料庫管理系統</p> <p>C 是，因為 JSON 是唯一可在 R 中存取和讀取的線上數據文件格式</p> <p>D 是，因為 JSON 僅僅是文字(text-only)檔案，並且可以由任何程式 語言所讀取</p>
5	假設數據分析師利用 R 程式語言進行工作，是否可以存取和分析全球綠色保險(GGI)核心資料庫的資訊？	<p>A 否，因為 R 可用的套件，僅僅是設計用於讀取線上數據集，而不是資料庫</p> <p>B 否，因為 R 不能與專屬資料庫(例如 GGI 資料庫)一起使用</p> <p>C 是，因為 R 提供了許多不同的套件來處理常見的資料庫類型，例如 MySQL</p> <p>D 是，因為 GGI 資料庫資訊是非結構化的</p>

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
提取與準備(Ingestion and Preparation)	1	C	6

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關數據檢查的問題。

在數據分析師開始數據清理或數據整理之前，他們首先需要查看他們正在處理什麼數據

請記住為每 1 項問題選擇 2 個答案。

1	從圖 2 中的資訊可以得出以下 2 項觀察結果？
	<ul style="list-style-type: none"> A 原始數據集包含結構化數據 B 原始數據集中沒有不相關、缺失或損壞的值 C 分析中包含 11 種保險產品 D 圖 2 中的資訊顯示在數據框中 E 每個目的地的定價結構(pricing structure)都不同
2	根據圖 2 中每個變量顯示的值，可以得出哪 2 項觀察結果？
	<ul style="list-style-type: none"> A “Age” 的類別中有 81 筆資料 B 該數據集包含有關 63,326 筆客戶的資訊 C 有 63,326 筆過時的數據(obsolete data entries) D 有 16 種不同的分銷渠道(distribution channels) E 保險在 149 個不同的銷售目的地(different destinations)
3	關於圖 2 所顯示的變量屬性(properties)，可以進行哪 2 項觀察？
	<ul style="list-style-type: none"> A Durations 顯示為numeric values B Gender 類型有兩種：“F” 和 “M” C Commission.in.value 的範圍從 9.57 到 11.88 D 每個 “Agency” 均使用唯一的三字母參考代碼記錄 E GGI product catalogue 包含 26 種不同的保險產品
4	圖 2 中哪 2 個變量以數值(numeric values)形式顯示？
	<ul style="list-style-type: none"> A Agency.Type B Duration C Net.Sales D Commission.in.value E Age
5	關於圖 3 的摘要統計數據中顯示的 “Gender” 資訊，可以做出以下 2 種觀察？
	<ul style="list-style-type: none"> A 總共向男性和女性出售了 45,107 種產品 B 有 45,107 筆損壞、不相關、丟失或未分類的數據。 C 購買保險產品的男性多於女性 D 無法確定購買保險產品的男性是否多於女性 E “Cancellation Plan” 是男性購買的最受歡迎的產品
6	從圖 3 的匯總統計可以得出以下 2 個觀察結果？
	<ul style="list-style-type: none"> A 最常購買保險的年齡是 39 歲 B “Travel Agencies” 銷售 GGI 保險產品的數量是 “Airlines” 的兩倍 C 購買保險產品的人群的四分位數為 35 歲 D 提出索賠的人數的百分比是為 9.27 E 最暢銷的保險產品是 “Cancellation Plan”

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
提取與準備(Ingestion and Preparation)	1	D	4

使用“情境手冊”為此問題提供的附加資訊，回答以下有關清理和整理數據的問題。

CMO 特別想知道哪一個代理商獲得了最高的佣金 (“Commission.in.value”)，以及客戶是男性還是女性。為了準備此項分析，數據分析師已在 R 中安裝了dplyr 套件，並將針對圖 2 和圖 3 中審查的數據集，使用常見的數據清理和整理操作。

為了對這些數據進行準確的分析，數據分析師還需要在不扭曲結果(without skewing the results)的情況下解決缺失的性別 (Gender) 值。

下表中的第 1 至 4 行由一個申論語句(assertion statement) 和一個原因語句(reason statement) 組成。對於每一行，請從選項 A 到 E 標識適用的適當選項。每個選項可以使用一次，使用超過一次，也可以不使用。

申論語句與原因語句的選項

1. 兩者都為真，原因語句也同時解釋了申論語句
2. 兩者都為真，但原因語句無法解釋申論語句
3. 申論語句正確與原因語句錯誤
4. 申論語句錯誤與原因語句正確
5. 申論語句錯誤與原因語句錯誤

	申論語句 (Assertion)		原因語句 (Reason)
1	僅顯示 “Agency”, “Claim”, “Destination”, “Commission.in.value” 和 “Gender” 列 數據分析師 應 使用 filter() 操作	因為 Because	Filter() 函數允許您根據指定的 任何條件選擇列 (select columns)
2	要按 “Commission.in.value” 列中的值 對數據進行排序 (從最高值開始 數據分析師 應使用 arranging() 函數。	因為 Because	arrange() 函數可以用來在不對 數據進行任何重大改變的情況 下，在列名 (column names) 中 產生 一致性。
3	為了消除所有 “Gender” 行 (rows) 的 缺失值 數據分析師應使用 drop_na() 函數。	因為 Because	drop_na() 函數使您可以消除任 何缺失值的數據行。
4	使用 fill() 函數，用上面一行 的值 來 替換每個缺少的 “Gender” 值可能會產生不同的結果。	因為 Because	Fill() 函數的結果將完全取決於行 (rows) 的排序方式。

題目編號	2
所涵蓋的課程大綱範圍	探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)	2	A	4

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關探索型數據分析 (exploratory data analysis) 的問題。

在嘗試進行任何正式建模之前，已要求數據分析師使用一些簡單的圖形來探索保險銷售 (Insurance Sales) 數據集。

請記住，為每項問題選擇 2 個答案

1	數據分析師在此探索階段應嘗試實現哪兩個目標？
	<p>A 更新並修改原始數據，以確保其對處理有用且有效</p> <p>B 確定是否有跡象表明向男性和女性銷售保險產品</p> <p>C 從數據集中檢測並更正或刪除不完整的客戶記錄</p> <p>D 重新格式化(reformat)數據集以提供最相關的模型構建結構</p> <p>E 更好地了解數據集中的變量</p>
2	從圖 4 的直方圖中可以得出以下 2 個觀察結果？
	<p>A 35 至 40 歲之間的女性是新保險產品的合適目標</p> <p>B 在 115 - 120 歲之間存在異常數據(outlier data)</p> <p>C 購買保險的女性年齡中位數為 60 歲</p> <p>D 最受歡迎的保險產品由 35 - 40 歲的女性購買</p> <p>E 婦女購買保險的平均年齡為 60 歲</p>
3	從圖 5 的箱型圖可以得出以下 2 個觀察結果？
	<p>A 購買保險產品的男性多於女性</p> <p>B 男性的中位數、全距和四分位數間距大於女性</p> <p>C 女性的平均年齡在 40 歲以下</p> <p>D 有針對 80 歲以上女性的異常值</p> <p>E 女性銷售的 75% 是針對 30 至 50 歲的女性</p>
4	從圖 6 的散點圖中可以得出以下 2 個觀察結果？
	<p>A “Age” 和 “Net.Sales” 變量之間存在負相關</p> <p>B “Age” 和 “Net.Sales” 變量之間存在正相關</p> <p>C 四分位數間距介於 0 到 85 之間</p> <p>D “Net.Sales” 值為 0 到 125 的產品比 “Net.Sales” 值 125 到 250 的產品多</p> <p>E 只有 8 或 9 個觀測值的 “Net.Sales” 值高於 225</p>

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)	2	B	6

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關回答以下有關統計推論 (statistical inference) 的問題。

1	定義“男性和女性在保險上的平均花費相同”的原假設(null hypothesis)的目的是什麼？
	<p>A 證明替代假設(alternative hypothesis)，即女性在保險上的平均支出高於男性</p> <p>B 反對原假設，即男女在保險上的平均支出是相同的</p> <p>C 證明男女在保險上的平均花費相同的原假設</p> <p>D 反對替代假設，即女性在保險上的平均支出高於男性</p>
2	圖 7 中觀察到的差異統計量(difference statistic)是多少？
	<p>A 3.3</p> <p>B 22.0</p> <p>C 25.3</p> <p>D 33.9</p>
3	計算觀察到的差異統計，為什麼數據分析師隨後使用“infer”套件產生 100 個排列(permutations)？
	<p>A 查看觀察到的差異統計是否僅僅是偶然的結果(result of chance alone)</p> <p>B 為 100 個隨機樣本中的每個樣本產生觀察到的差異統計量</p> <p>C 假設原假設不是真的情況下，產生 100 個隨機樣本</p> <p>D 查看計算出的差異統計量是否僅僅是偶然的結果(result of chance alone)</p>
4	使用行業標準閾值拒絕原假設，對保險數據隨機樣本的 100 個排列結果進行分析可以得出什麼結論？
	<p>A 沒有足夠的統計證據來拒絕原假設(null hypothesis)</p> <p>B 現在已經證明了替代假設(alternative hypothesis)，因此應該拒絕 原假設(null hypothesis)</p> <p>C 原假設(null hypothesis)應被拒絕，因為替代假設是可行的</p> <p>D 由於原假設(null hypothesis)是可行的，因此應拒絕替代假設</p>
5	哪種圖形適合於直觀地表示 100 個數據排列的結果？
	<p>A 線條圖，針對第 5 個百分點和第 95 個百分點繪製每個排列</p> <p>B 線條圖，將觀察到的每個排列的差異統計與計算的差異統計作圖</p> <p>C 直方圖，顯示了 100 個不同排列與觀察到的差異統計量進行比較</p> <p>D 直方圖，顯示觀察到的差異統計與其他觀察到的差異的比較</p>
6	為什麼統計推論(statistical inference)的流程對於分析很重要？
	<p>A 降低保險銷售數據集中的數據可能會干擾假設檢定的有效性的風險</p> <p>B 估計假設檢定中，沒有有效結果的不確定性</p> <p>C 可以根據該母體子集得出有關潛在 GGI 客戶購買行為的結論</p> <p>D 預測保險銷售數據集中的偏差水平，該偏差水平可能會影響假設檢定的結果</p>

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)	2	C	4

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關回答以下有關統計關聯 (**correlation**) 的問題。

1	<p>“Duration” 和 “Net.Sales” 之間似乎存在很強的線性關聯性。這是對圖 8 之散點圖的適當評估嗎？</p> <p>A 否，因為通過的數據點沒有顯示清晰的線性</p> <p>B 否，因為散點圖無法指示兩個變量之間是否可能有任何關係</p> <p>C 是，因為大多數數據點的持續時間都低於 100 個月</p> <p>D 是，因為 “Duration” 和 “Net.Sales” 之間存在明確的關係</p>
2	<p>皮爾遜關聯係數的值 0.3143053 是否表示 “Duration” 與 “Net.Sales” 之間，呈正相關？</p> <p>A 否，因為皮爾遜關聯係數不是用來表示變量之間的關係的。</p> <p>B 否，因為小於 0.5 的值沒有意義，應該忽略</p> <p>C 是，因為介於 0 和 +1 之間的值表示正線性相關</p> <p>D 是，因為此值表示 “Duration” 的增加將 “Net.Sales” 的減少</p>
3	<p>數據分析師認為，皮爾遜關聯係數的值證明 “Net.Sales” 是由保險產品的 “Duration” 引起的。這是正確的觀察嗎？</p> <p>A 不，因為當變量的維度不同時，就沒有因果關係</p> <p>B 不，因為存在關聯性，並不意味著所購買產品的 “Duration” 是起因 (is caused by) 於 “Net.Sales” 引起的</p> <p>C 是，因為相關值顯示 “Net.Sales” 和 “Duration” 朝相反方向移動 的程度</p> <p>D 是，因為 “Duration” 和 “Net.Sales” 的值一起增加或減少</p>
4	<p>數據分析師認為，相應的關聯係數 0.3143053 表明，隨著 “Duration” 的增加， “Net.Sales” 值可能會增加。這是現實的期望嗎？</p> <p>A 否，因為該關聯係數表明一個變量方向的變化對另一個變量沒有影響</p> <p>B 否，因為該關聯係數描述了這兩個變量在相反方向上移動的程度</p> <p>C 是，因為此關聯係數意味著一個變量導致另一個變量移動</p> <p>D 是，因為此關聯係數意味著這兩個變量的值一起增加或減少</p>

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
探索型數據分析、推論、關聯和回歸 (Inference, Correlation and Regression)	2	D	6

回答以下有關簡單線性回歸 (simple linear regression) 的問題

首席產品官已要求數據分析團隊幫助調查保險產品的價格是否會影響所銷售 保單的數量。他希望能夠預測最佳價格點，以使新保險產品的銷售數量最大 化。

數據分析團隊基於以下假設創建了回歸模型：給定產品的銷售數量（“Product.Volume”）取決於產品價格（“Product.Price”）。

Call:

```
lm(formula = Product.Volume ~ Product.Price, data = Insurance_Data)
```

下表中的第 1 至 6 行由一個申論語句(assertion statement) 和一個原因語句(reason statement) 組成。對於每一行，請從選項 A 到 E 標識適用的適當選項。每個選項可以使用一次，使用超過一次，也可以不使用。

申論語句與原因語句的選項

1. 兩者都為真，原因語句也同時解釋了申論語句
2. 兩者都為真，但原因語句無法解釋申論語句
3. 申論語句正確與原因語句錯誤
4. 申論語句錯誤與原因語句正確
5. 申論語句錯誤與原因語句錯誤

	申論語句 (Assertion)		原因語句 (Reason)
1	簡單線性回歸是合適於此分 析的模型	因為 Because	在簡單的線性回歸中，單一個 自變量用於預測因變量的值
2	簡單的線性回歸應該顯示出變更 “Product_Price” 的值，將 如何影響 “Product_Volume”	因為 Because	簡單的線性回歸影響因變量， 以優化自變量的結果。
3	如果線性回歸模型顯示價格上漲會影響銷售數量的減少，那麼皮爾遜關聯係數將 為負	因為 Because	關係的強度由斜率的值及其與 +1 或-1 的接近程度來指示
4	如 果 “Product_Volume” 和 “Product_Price” 之間存在負相關關聯，則最佳擬合線應 在該線下方顯示大量數據點	因為 Because	最佳擬合線是簡單的線性回歸線，可將平方誤差的總和最小化
5	最佳擬合線性回歸線將在給定特定產品價格的情況下預測產品銷售數量	因為 Because	“最佳擬合” 是用於描述線性回歸線與 y 軸交叉的點的術語
6	模型的準確性將由殘差值指示	因為 Because	殘差平方的總和表示預測值與實際值的接近程度

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
分類模型 (Classification Models)	3	A	5

使用“情境手冊”中針對此問題提供的附加資訊，回答有關 kNN 運作 (k-Nearest Neighbour operation) 的以下問題。

1	在建構 k-NN 模型時，會利用測試數據產生多少個預測？
A	16,630
B	3,326
C	60,000
D	63,326
2	數據分析師在第一個 k-NN 模型中使用的 k 值是多少？
A	0
B	0.4
C	0.45857
D	1
3	第一個 k-NN 模型對觀測值的分類正確率不到 50%。數據分析師可以做什麼來嘗試並提高模型的準確性？
A	再次運行該模型，但從訓練數據集中排除任何常態化的數據
B	減少訓練數據集中的數據點數量
C	用不同數量的最近鄰進行實驗，以找到 k 的最佳值
D	使用不同的訓練數據集和測試數據集之間的分割來運行模型
4	數據集中有兩個變量是數字變量，有不同的維度和範圍。這兩個變量是 “Age” 和 “Average Family Income”。變量 “Age” 的範圍從 0 到 110，變量 “Average Family Income” 的範圍從 0 到 80 萬。這對模型的準確性 可能有什麼影響？
A	沒有，因為這兩個變量都是數字
B	沒有，因為這兩個變量的範圍都是從 0 開始的
C	影響很小，因為數據集中的其他三個變量都是非數字型的
D	“Average Family Income” 很可能對結果產生不成比例的影響
5	使用顯示多個 k-NN 操作結果的表格，哪種說法是對模型精度是正確解釋？
A	使用 k 的那一個值並不重要，因為精度值都小於 1
B	當 k 的值從 3 增加到 7 時，模型的精度會降低
C	k=7 的值將產生最準確的模型
D	k=3 的值將返回最低的錯誤數 (lowest number of errors)

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
分類模型 (Classification Models)	3	B	5

使用“情境手冊”中針對此問題提供的附加資訊，回答以下有關單純貝葉斯分類器 (Naïve Bayes Classifier) 應用的問題。

1	<p>哪種說法適用於單純貝葉斯模型？</p> <p>A 如果首先將數據常態化(標準化)，則產生最準確的結果</p> <p>B 使用只有一個或兩個變量的數據集時，效果最佳</p> <p>C 能夠進行快速計算，因此適合即時使用</p> <p>D 是解決推論型業務目標(inferential business objectives)的關鍵技術</p>
2	<p>圖 9 所示的概率表缺少哪些資訊？</p> <p>A 混淆矩陣(confusion matrix)，顯示 “Age” 和 “Duration” 的平均值</p> <p>B 一系列的維恩圖，顯示了條件概率的各種排列</p> <p>C 每個 “Gender” 和 “Product” 的組合以及每個 “Age” 與 “Product” 的組合之條件概率(Conditional probabilities)</p> <p>D 顯示預測類與實際類表格對比的部分</p>
3	<p>如果條件聯合概率之一的值為零，數據分析師應採取什麼措施？</p> <p>A 指定將此聯合概率從模型中排除</p> <p>B 將 laplace 參數= 1 添加到 naïve_bayes() 函數</p> <p>C 清除數據以刪除所有零值</p> <p>D 將 naïve_bayes() 函數的結果添加到原始數據集中，然後再次運行該演算法</p>
4	<p>數據分析師正在考慮是否將 “Location” (客戶的地理屬性)也包括在模型中。什麼陳述能最好地說明這一點的可能含義？</p> <p>A 不建議更改，因為同時包括人口統計和地理變量將降低模型的預測能力</p> <p>B “Location” 不能包括在內，因為該模型不適用於地理和人口變量</p> <p>C 由於所有變量都是獨立的，因此添加客戶 “Location” 將進一步提高模型的預測複雜性</p> <p>D 因為模型只能使用最多兩個變量，所以需要替換變量 “Gender” 或 “Age”</p>
5	<p>數據分析師應如何計算此單純貝葉斯模型的準確性？</p> <p>A 將 laplace 參數= 0 添加到 naïve_bayes() 函數</p> <p>B 比較條件概率和先驗概率</p> <p>C 重複執行模型多次，然後比較結果</p> <p>D 比較整個數據集的實際類別和預測類別</p>

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
分類模型 (Classification Models)	3	C	6

使用“情境手冊”中為此問題提供的附加資訊，回答以下有關邏輯回歸 (logistic regression) 的問題。

下表中的第 1 至 6 行由一個申論語句(assertion statement) 和一個原因語句(reason statement) 組成。對於每一行，請從選項 A 到 E 標識適用的適當選項。每個選項可以使用一次，使用超過一次，也可以不使用。

申論語句與原因語句的選項

1. 兩者都為真，原因語句也同時解釋了申論語句

2. 兩者都為真，但原因語句無法解釋申論語句

3. 申論語句正確與原因語句錯誤

4. 申論語句錯誤與原因語句正確

5. 申論語句錯誤與原因語句錯誤

	申論語句 (Assertion)		原因語句 (Reason)
1	數據集 Claim_Data 中的所有輸入變量的值都必須為“是” 或 “否”	因為 Because	邏輯回歸模型中的目標變量有兩個可能的結果
2	此數據集感興趣的二進制變量為 “Cliam”	因為 Because	邏輯回歸模型的輸入變量可以是分類變量 (categorical variables)
3	為了預估未來的行為，將預測操作應用於此邏輯回歸模 型是適當的	因為 Because	邏輯回歸將現有變量之間的關係轉換為概率，它不是分類 器。
4	對於 GGI，將超過 0.2189 閾值的客戶分類為潛在索賠人 是適當的	因為 Because	理解樣本數據集中，實際分類的概率分數將影響你對未來預 測的分類設置的閾值。
5	邏輯回歸模型的準確率為 98.5%	因為 Because	閾值設置將影響模型的準確性
6	設置一個閾值，使預測類的數量將高於觀察到的分類 將使模型更加準確	因為 Because	模型的準確度計算為 100%減去為閾值設置的值

EBDA-CH-AX01-V.

Page 14 of 20
©APM Group Ltd 2021.

Document Owner - Chief Examiner

未经APM Group Ltd. 明确许可，不得重新制作或转售本文件。

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
分類模型 (Classification Models)	3	D	4

使用“情境手冊”中為此問題提供的附加資訊，回答以下有關構建分類樹 (classification tree) 演算法的問題。

對於以下的情況，請確定所採取的措施是否適當，然後選擇支持您的決定選項。

1	在執行第一個模型之前，數據分析師利用從具有最多類別的數據集中選擇變量來確定演算法的第一個拆分標準，這是數據分析師採取的適當步驟嗎？	
	A	否，因為數據分析師應該已經為第一個拆分條件標識了一個二進制變量
	B	否，因為該演算法基於將產生最具相同性質子組的結果來確定第一個拆分標準
	C	是，因為如果將最多類別的變量用於第一個拆分條件，則存在更多的同類子組
	D	是，因為第一個拆分標準是必須在演算法中指定的第一個參數
2	在設計第一個模型時，數據分析師為 maxdepth 設置一個值。這是數據分析師設置的適當參數嗎？	
	A	否，因為當明確指定輸出變量中的類別數量時，不需要 maxdepth
	B	否，因為應該設計分類樹以完全適合訓練數據集
	C	是，因為預修剪樹將有助於管理最終結果的大小和深度
	D	是，因為數據集中的 11 個變量需要修剪
3	產生未修剪的決策樹後，數據分析師使用 Cp 圖（複雜性參數）的結果來提高模型的準確性。這是數據分析師採取的適當措施嗎？	
	A	否，因為輸入變量的數量接近 Cp 圖中所示的最佳樹大小
	B	否，因為 Cp 圖沒有顯示樹的任何最佳大小
	C	是，因為 Cp 圖顯示樹中有太多分支
	D	是，因為 Cp 圖表明複雜度參數在 0.0047 附近可提供最佳精度
4	數據分析師告訴執行團隊，最多，分類樹演算法可以根據客戶的特徵將 82.9% 的客戶分類為正確的類別。這是數據分析師採取的適當措施嗎？	
	A	否，因為 Cp 圖顯示有一些 Cp 值可用於改善樹的形狀和準確性
	B	否，因為 Cp 圖顯示的精度為 95.3%
	C	是，因為已使用測試數據集確定準確性
	D	是，因為這是修剪後預測操作的平均值

題目編號	4
所涵蓋的課程大綱範圍	聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)	4	A	6

回答有關層次聚類 (hierarchical clustering) 的以下問題。

數據分析團隊正在嘗試不同的模型，以了解如何利用 GGI 資料庫來根據客戶的特徵識別不同的客戶群。客戶資料庫中的以下變量已經導入到名為

“Customer Profiles” 的數據集中：

- Age (年為單位)
- Duration (保險單時間 (以月為單位))
- Spend (貨幣為單位)

請記住，為每項問題選擇 2 個答案。

1	層次聚類模型的哪 2 個特徵，使其適合使用於此分析？
	<p>A 無需指定輸出組將是什麼</p> <p>B 層次聚類模型僅適用於順序數據 (ordinal data)</p> <p>C 該模型將從一組標記的樣本數據中學習</p> <p>D 輸出不受數據中缺失值的影響</p> <p>E 該模型可用於評估多個變量</p>
2	哪 2 項操作說明了層次聚類模型將如何分組 GGI 客戶？
	<p>A 它使用樹狀圖 (dendrogram) 來確定每位客戶之間的任何依存關係</p> <p>B 迭代 (iteratively merges) 的，將最相似的客戶群合併在一起</p> <p>C 使用集群大小的寬度和高度參數為客戶標識適當的集群</p> <p>D 將所有客戶歸為一個組，然後依次將客戶分為不同的組</p> <p>E 它測量客戶數據觀測值之間的距離</p>
3	數據分析師在構建此案例的層次聚類模型時需要使用哪種 2 種技巧？
	<p>A 手肘法 (Elbow method)</p> <p>B 歐幾里得距離 (Euclidian distance)</p> <p>C Jaccard 指數 (Jaccard Index)</p> <p>D 標準化 (Standardization)</p> <p>E 推論 (Inference)</p>

請翻頁繼續回答

4	數據分析師正在決定是使用完全鏈接或是平均鏈接 (Complete Linkage or Average Linkage) 的方法，將數據點分配給集群。關於這兩種方法，哪 2 項陳述是正確的觀察結果？
A	如果使用“平均鏈接”方法，則集群將形成不同的模式
B	使用“平均鏈接”方法，產生的 B 集群更有可能受到異常值(outliers) 的影響
C	如果使用“完全鏈接”方法，該模型可能會更好地檢測細微模式
D	如果必須以數值形式顯示數據，則這兩種方法都不適用
E	如果預先確定 cluster 數，則所用的鏈接方法將無關緊要
5	哪 2 項操作將決定模型產生客戶資料或分組的數量與它們的方式？
A	分配最大深度 (Assign the maximum depth)
B	指定 k 的值 (Specify the value of k)
C	定義最大高度 (Define the maximum height)
D	確定 n 的大小 (Determine the size of n)
E	計算最小分裂 minsplitt (Calculate the minsplitt)
6	在解釋此聚類活動的最終結果時，數據分析師應進一步調查哪些 2 個 觀察結果？
A	所有變量，均具有較低平均值的集群
B	客戶數量少的集群中，變量的極端值
C	標準差為 1 的任何變量
D	平均值為 0 的任何變量
E	僅包含一個或兩個客戶的集群

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)	4	B	5

回答以下有關 k 均值聚類 (k-means clustering) 的問題。

數據分析人員希望了解使用相同樣本數據集時，k 均值聚類模型的結果與分層聚類模型的結果有何不同。客戶檔案樣本數據集包含每位客戶的以下資訊：

- Age
- Duration of insurance policy (保險單時間)
- Spend on insurance policy (花費在保單上)

1	為什麼數據分析師必須標準化 (standardize) “客戶資料” 數據集？
A	“Age”，“Duration” and “Spend” 將分別具有不同的維度單位 (different dimensions)
B	對於這些變量，將需要使用 Jaccard 指數
C	所有這三個變量的平均值必須為 1
D	使用這些變量建立模型將花費很長時間
2	k 均值模型的聚類方法與分層聚類演算法的聚類方法有何不同？
A	只有 k 均值模型將繼續對數據點進行分組，直到剩下一個聚類為止
B	可能會產生更少的聚類，並且它們應該更靠近在一起
C	繼續的向聚類分配數據點，直到聚類中沒有任何更改為止
D	取決於是否指定相同或不同的鏈接方法
3	以下哪個是 k-means 模型執行的操作之一？
A	將每個觀測值分配給最近的質心 (closest centroid)
B	在演算法的第一次迭代後，固定質心的位置
C	使用高度作為切入標準，確定質心的最終位置
D	用兩個最同質的變量的平均值，確定 “假設” 的質心
4	數據分析師被要求對具有最大相似性的客戶進行分組，同時使每個組之間達到最佳的不相似性。數據分析師應該怎樣做才能確定他們應該 針對多少個客戶群？
A	信任 k 均值演算法，可確定最佳資料文件的數量
B	使用手肘法進行實驗以找到 k 的值
C	詢問市場經理他們希望有多少個人資料
D	在組合相似的群集之前，指定群集之間允許的最小距離
5	將手肘法與 k-means 演算法一起使用時，數據分析師正在尋找的關鍵 資訊是什麼？
A	平均群集大小突然減小的點
B	繪製 k 時的傾斜角度與平方誤差之和
C	平方誤差之和開始趨於平緩的 k 值
D	任何群組的邊界 “壓迫” 其他群組邊界的地方

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)	4	C	4

回答以下有關異常質檢測 (Outlier Detection) 的問題。

數據分析師正在使用 Grubbs 檢驗來發現，向代理商銷售 GGI 保險產品 (“Commission.in.value”) 的佣金 (“Net.Sales”) 是否明顯高於或低於其他佣金。

下表中的第 1 至 4 行由一個申論語句(assertion statement) 和一個原因語句(reason statement) 組成。對於每一行，請從選項 A 到 E 標識適用的適當選項。每個選項可以使用一次，使用超過一次，也可以不使用。

申論語句與原因語句的選項

1. 兩者都為真，原因語句也同時解釋了申論語句
2. 兩者都為真，但原因語句無法解釋申論語句
3. 申論語句正確與原因語句錯誤
4. 申論語句錯誤與原因語句正確
5. 申論語句錯誤與原因語句錯誤

	申論語句 (Assertion)		原因語句 (Reason)
1	數據分析師應首先明確一個佣金支付需要與其他佣金支付有多大的不同，才能被認為是異常值。	因為 Because	Grubbs 檢定在數據隨機分佈的假設下工作
2	在 Grubbs 檢定中，概率為 0.333 的數據點不太可能是一個離群點	因為 Because	Grubbs 檢定評估 p-value 是否為 0 到 1 之間的數字
3	為了使 Grubbs 檢定有效運行，數據分析師在執行測試時將需要選擇 “Net.Sales” 和 “Commission.in.value”	因為 Because	至少需要兩個變量才能探索直方圖中數據集的分佈
4	Grubbs 檢定將把與其他佣金支付的歐幾里得距離最大的佣金支付識別為最大離群值	因為 Because	出現在散點圖或直方圖最右側的孤立數據點很可能是異常值

所涵蓋的課程大綱範圍	題目編號	子題編號 (Part)	分數 (Mark)
聚類、異常值檢測與可視化 (Clustering, Outlier Detection & Visualization)	4	D	5

回答以下有關數據呈現 (data presentation) 的問題。

對於以下每個數據呈現活動，請確定所採取的措施是否適當，然後選擇支持您的決定的選項。

1	<p>大數據卓越中心整理了一個 Codebook，記錄了每次分析中使用的數據文件的內容、結構和佈局，以及如何解釋這些數據。儘管有多個人可以更改代碼本中的資訊，但是該資訊已被正式記錄並受版本控制。這是共享有關分析資訊的適當方法嗎？</p> <p>A 否，因為對記錄的有關分析的資訊，進行的任何更改都會使它對將來分析的有用性失效</p> <p>B 否，因為一旦分析完成，此資訊將不再具有任何意義</p> <p>C 是，因為考慮到每次分析所花費的時間至關重要</p> <p>D 是，因為這將使其他人將來可以復制數據分析</p>
2	<p>數據分析師正在填寫 InsuranceSales 數據集的 Codebook，並正在考慮記錄有關遺失數據 (missing data) 的資訊。他決定記錄每個變量的缺失值 總數就足夠了。這是否為描述本 Codebook 中的缺失值的適當作法？</p> <p>A 否，因為資訊應包括缺失值在數據集中的顯示方式</p> <p>B 否，因為應該從數據集中刪除缺少的值</p> <p>C 是，因為獲得此資訊非常重要，因為丟失的數據可能會使分析產生偏差</p> <p>D 是，因為必須計算任何值的 “N/A” 與 “0” 的資料數</p>
3	<p>首席執行官提醒首席數據官，GGI 是一個全球性組織，其客戶遍布五大洲。為了解決不同受眾的語言和文化差異，首席執行官建議，應採用針對每個特定受眾設計的格式來呈現數據分析的可視化。這是確保所有人都了解圖形的適當方法嗎？</p> <p>A 不，因為全世界的每個人對可視化的解釋都是相同的</p> <p>B 不，因為數據分析的結果與每個人都不相關</p> <p>C 是，因為不同文化的人對可視化的理解不同</p> <p>D 是，因為圖形數據的顯示將需要符合國家標準</p>
4	<p>一位數據分析師試圖顯示數據集中三個變量之間的關係，每個變量具有不同的維度和範圍。這些變量之一是二進制的，並且會影響其他兩個變量。數據分析師選擇使用顏色顯示散點圖。這是呈現此數據分析結果的合適方法嗎？</p> <p>A 否，因為散點圖具有二維坐標，即 x 軸和 y 軸</p> <p>B 否，因為散點圖的美學僅限於簡單的單變量分析</p> <p>C 是，因為散點圖的美學可以包括將二進制變量顯示為三維的顏色</p> <p>D 是，因為散點圖為快速比較多組數據點提供了機會</p>
5	<p>數據分析師希望使用 ggplot 從同一數據集創建兩個不同的可視化效果：條形圖和直方圖。為此，他將更改美學參數。這是實現此結果的合適方法嗎？</p> <p>A 否，因為數據不會有任何根本變化</p> <p>B 否，因為更改圖形的圖形形狀與幾何有關</p> <p>C 是，因為美學顯示變量如何映射到繪圖的幾何圖形</p> <p>D 是，因為幾何形狀相同，所以只有美學會改變</p>



企業大數據分析師考試

AX01

評分方案

注意：對於多選 (Multiple Response, MR) 題時，必須要選擇了所有正確答案時，才可得 1 分，否則僅得 0 分。

Question	Part	Type	Response	A	B	C	D	E	F	G	H	I
1 (IP)	A	MG	1	0	0	0	0	1	0			
			2	0	1	0	0	0	0			
			3	0	0	1	0	0	0			
			4	1	0	0	0	0	0			
			5	0	0	0	0	0	1			
1 (IP)	B	CL	1	1	0	0	0					
			2	0	0	0	1					
			3	0	0	1	0					
			4	0	0	0	1					
			5	0	0	1	0					
1 (IP)	C	MR	1	1	0	0	1	0				
			2	0	1	0	0	1				
			3	0	0	0	1	1				
			4	0	0	1	1	0				
			5	0	1	0	1	0				
			6	0	1	0	0	1				
1 (IP)	D	AR	1	0	0	0	0	1				
			2	0	0	1	0	0				
			3	0	0	0	1	0				
			4	1	0	0	0	0				

Question	Part	Type	Response	A	B	C	D	E	F	G	H	I
2 (ER)	A	MR	1	0	1	0	0	1				
			2	1	1	0	0	0				
			3	0	1	0	1	0				
			4	0	0	0	1	1				
2 (ER)	B	CL	1	0	1	0	0					
			2	1	0	0	0					
			3	1	0	0	0					
			4	0	0	1	0					
			5	0	0	1	0					
			6	0	0	1	0					
2 (ER)	C	CL	1	1	0	0	0					
			2	0	0	1	0					
			3	0	1	0	0					
			4	0	0	0	1					
2 (ER)	D	AR	1	1	0	0	0	0				
			2	0	0	1	0	0				
			3	0	1	0	0	0				
			4	0	0	0	1	0				
			5	0	0	1	0	0				
			6	0	1	0	0	0				

Question	Part	Type	Response	A	B	C	D	E	F	G	H	I
3 (CM)	A	CL	1	0	1	0	0					
			2	0	0	0	1					
			3	0	0	1	0					
			4	0	0	0	1					
			5	0	0	1	0					
3 (CM)	B	CL	1	0	0	1	0					
			2	0	0	1	0					
			3	0	1	0	0					
			4	0	0	1	0					
			5	0	0	0	1					
3 (CM)	C	AR	1	0	0	0	1	0				
			2	0	1	0	0	0				
			3	1	0	0	0	0				
			4	1	0	0	0	0				
			5	0	0	0	1	0				
			6	0	0	0	0	1				
3 (CM)	D	CL	1	0	1	0	0					
			2	0	0	1	0					
			3	0	0	0	1					
			4	1	0	0	0					

Question	Part	Type	Response	A	B	C	D	E	F	G	H	I
4 (CV)	A	MR	1	1	0	0	0	1				
			2	0	1	0	0	1				
			3	0	1	0	1	0				
			4	1	0	1	0	0				
			5	0	1	1	0	0				
			6	0	1	0	0	1				
4 (CV)	B	CL	1	1	0	0	0					
			2	0	0	1	0					
			3	1	0	0	0					
			4	0	1	0	0					
			5	0	0	1	0					
4 (CV)	C	AR	1	0	0	0	1	0				
			2	0	1	0	0	0				
			3	0	0	0	0	1				
			4	0	0	0	1	0				
4 (CV)	D	CL	1	0	0	0	1					
			2	1	0	0	0					
			3	1	0	0	0					
			4	0	0	1	0					
			5	0	1	0	0					



ENTERPRISE BIG DATA ANALYST

企業大數據分析師考試

AX01

解答 (Rationale)

Question: 1, Syllabus: IP, Part: A, Type: MG, SyllabusRef: IP0301, Level: 3		
1	Correct [E]:	因果型業務目標(Causal business objective)是為了理解為什麼會發生某種現象。因果型業務目標在回答改變一個屬性是否會改變另一個屬性。因果型分析的目的是嘗試查找問題的根本原因，而不是查找症狀。因果型分析技術有助於發現導致某種情況的事實。(請參閱第29頁)
2	Correct [B]:	探索型業務目標(An exploratory business objective) 是為了發現(未知)模式、趨勢或兩個或多個變量之間的關係。探索型練習的結果很可能是沒有發現任何結果。由於存在這種不確定性，探索型數據分析通常也稱為“數據挖掘”(data mining)。(參考第25頁)
3	Correct [C]:	推論型業務模式(An inferential business objective)的目標是利用檢查母體的子集(樣本)來找到母體的概括。推論型業務目標可以與描述型業務目標進行對比。描述型的業務目標僅與觀察到的數據的屬性有關，並且不會基於數據來自較大母體的假設。(參考第26頁)
4	Correct [A]:	描述型業務目標(A descriptive business objective)是總結數據集的特徵。描述型業務目標試圖找到一個特定的事實，並且不接受解釋。描述型業務目標只能有一個正確答案。(請參閱第24頁)
5	Correct [F]:	機理型業務目標(Mechanistic business objectives)提供了業務問題的答案，這些業務問題可以用純粹的實際性或確定性術語來解釋現象。機理業務目標試圖找到原因的答案，並且比因果業務目標更進一步。因果業務目標僅僅是確定一個變量是否直接導致另一變量，而機理型目標是可以解釋這種關係的原因。(請參閱第29頁)

Question: 1, Syllabus: IP, Part: B, Type: CL, SyllabusRef: IP0402, Level: 4			
1	A	Correct:	為了將CSV檔案導入和讀取到R中，可以使用read.csv() 指令。但是，當數據集的第一行包含標頭時，標頭參數應設置為“TRUE”。在問題中，標頭參數設置為“FALSE”。(參考第33頁)
	B	Incorrect:	請參閱答案 A。公司資料庫的格式與此導入無關。
	C	Incorrect:	雖然 read.csv() 是最好用於讀取 R 中的 CSV 檔案，但命令不正確，且標頭參數應設置為“TRUE”。在問題中，標頭參數設置為“FALSE”。(參考第33頁)
	D	Incorrect:	請參見答案 A。使用 R 套件讀取其他類型的數據格式。(參考第33頁)
2	A	Incorrect:	該命令是合適的。參數“skip”的設置是確定是否要讀取標頭。(參考第35頁)
	B	Incorrect:	該命令正確，並且 read.lines 是用於導入 .txt 檔案的 R 命令。(參考第35頁)
	C	Incorrect:	請參閱答案D
	D	Correct:	導入文字檔案時，您需要指定文字檔案是否包含標頭。在這種情況下，標頭不存在，因此不需要跳過第一行，因此將參數“skip”設置為 0。(請參閱第35頁)
3	A	Incorrect:	Friends4Ever 數據確實包含標記數據(include markup data)，而 download.file() 方法是合適的方法。請參閱答案C
	B	Incorrect:	“Friends4Ever”是XML檔案，而download.file() 方法是適當的方法。請參閱答案C
	C	Correct:	download.file() 方法可用於從 Internet 下載所有類型的檔案。但是，某些網站(尤其是更高等的網站)確實設置了某些限制，阻止您下載他們的數據。在大多數情況下，這些網站將提供 API 來控制你的讀取使用。(請參閱第37頁)
	D	Incorrect:	這是一個合適的方法，因為該函數可以用來從 Internet 上下載所有類型的數據。雖然API經常被用來控制對數據的讀取使用，但這並不是download() 函數適合使用的原因。(請參閱第37頁)
4	A	Incorrect:	只有 XML 格式的檔案，才分為標記和內容(markup and content)。請參閱答案D
	B	Incorrect:	MySQL是一個開放源碼的關聯式資料庫管理系統。見答案D
	C	Incorrect:	R 中有許多功能可以導入和讀取不同格式的數據(線上和本地)。(請參閱第33–43 頁)。請參閱答案D。
	D	Correct:	JSON 提供一種文字格式，利用該格式可以輕鬆儲存和交換大量數據。最常見的用途是在 Web 瀏覽器和伺服器之間交換數據。由於 JSON 格式僅是文字，因此任何程式語言都可以輕鬆讀取和使用它。(參考第39頁)
5	A	Incorrect:	R 適合分析 GGI 數據，R可以與線上數據集和資料庫中的數據一起使用。(參考第33~43頁)
	B	Incorrect:	R 適用於分析GGI數據，因為它是MySQL資料庫(非專屬資料庫 proprietary databases)，並且有 R 套件可使用於此。也有 R套件可用於專屬資料庫。(參考第41頁)
	C	Correct:	R 適用於分析 GGI 數據，因為它是 MySQL 資料庫，並且為此提供了R 套件
	D	Incorrect:	每個 MySQL 資料庫均由許多表所組成，每張表均由許多行(稱為記錄)和列(稱為欄位)組成。它是結構化數據。(請參閱第41頁)。

Question: 1, Syllabus: IP, Part: C, Type: MR, SyllabusRef: IP0303, Level: 3			
1	A	Correct:	數據框(data frame) 是表或二維陣列狀結構，其中每一列(each column)各包含一個變量，而每一行(each row)包含來自每一列的一組值。這意味著您正在查看結構化數據。(請參閱第49頁)
	B	Incorrect:	無法從圖2的數據框中觀察到這一點，因為 str() 僅為您提供數據集核心屬性的一般概述。(請參閱第49頁)
	C	Incorrect:	圖2與數據框中的變量數量有關。該框架由11個列(columns) (11種變量) 和 63,326 行(rows) (63326 個觀察值)所組成。第4列(columns) (第4個變量) 表明有26種不同的產品名稱。(附加資訊)
	D	Correct:	數據框是表或二維陣列狀結構，其中每一列(each column)各包含一個變量，而每一行(each row)包含來自每一列的一組值。(請參閱第49頁)
	E	Incorrect:	從圖2所示的數據集的 structure() 中無法觀察到這一點，因為 str() 僅為您提供數據集核心屬性的一般概述。(請參閱第49頁)
2	A	Incorrect:	我們可以看到此列中的第一個數據點是81，它們是整數值(integer values)。81不代表此數據集中的數值數量。structure() 函數不會顯示列或變量中有多少個值。(請參閱第49頁)
	B	Correct:	數據框由11列(columns)(變量variables)和63,326行(rows)(觀察值observations)組成。(附加資訊)
	C	Incorrect:	有63,326個觀測值。缺少的數據顯示為""或N/A。見答案B。
	D	Incorrect:	存在2個不同的分發管道：離線或聯線。(附加資訊)
	E	Correct:	目的地列顯示了149個級別的因子(a factor of 149 levels)。因子(Factors)是用於對數據進行分類並將其存儲為級別的數據對象。它們既可以儲存字串，也可以儲存整數。它們對於有限數量的唯一值(如國家代碼)的列中很有用，因此它們可以很容易地進行排序。(請參閱第49頁)。
3	A	Incorrect:	這些是整數(integers)。數值為十進位制(decimals)。(請參閱第49頁)
	B	Incorrect:	顯示了3個不同的級別(3 different levels)："" (空白)、“F”、“M”。見答案D。
	C	Incorrect:	結構數據框(structure data frame)中，僅顯示前10個值。這並不代表完整的數值範圍。見答案D。
	D	Correct:	“Agency”列顯示16級因子(shows a factor of 16 levels)。因子(Factors)是用於對數據進行分類並將其存儲為級別的數據對象。它們既可以儲存字串，也可以儲存整數。它們對於有限數量的唯一值(如國家代碼)的列中很有用，因此它們可以很容易地進行排序。(請參閱第49頁)。
	E	Correct:	“Product Name”列顯示26級因子(a factor of 26 levels)。因子(Factors)是用於對數據進行分類並將其存儲為級別的數據對象。它們既可以儲存字串，也可以儲存整數。它們對於有限數量的唯一值(如國家代碼)的列中很有用，因此它們可以很容易地進行排序。(請參閱第49頁)。
4	A	Incorrect:	“Agency.Type”數據顯示為因子(shown as a Factor)。因子是用於對數據進行分類並將其存儲為級別的數據對象。它們既可以儲存字串，也可以儲存整數。它們對於有限數量的唯一值(如國家代碼)的列中很有用，因此它們可以很容易地進行排序。
	B	Incorrect:	“Duration”數據顯示為整數(Integer)。整數是可以為正、負或零的整數(不是小數 not a fractional number)。(請參閱第49頁)
	C	Correct:	“Net. Sales”顯示為數值(Numeric values)：十進位數在R中稱為數值。(請參閱第49頁)
	D	Correct:	“Commission.in.value”顯示為數值(Numeric values)：十進位數在R中稱為數值。(參見第49頁)。
	E	Incorrect:	“Age”顯示為整數(Integer)。整數是可以為正、負或零的整數(不是小數 not a fractional number)。(請參閱第49頁)
5	A	Incorrect:	45,107 顯示為N/A。見答案B。

	B	Correct:	45,107人顯示為“N/A”的性別因子(factor)。您處理的大多數數據集都不是完美的。數據集可能包含損壞、不相關或丟失的數據。缺失值(用NA或“”表示)是數據集中最常見的“缺陷 flaws”之一，它們之所以重要，是因為它們可能會對您要使用的函數和公式產生重大影響。(參考第61頁)
	C	Incorrect:	有45,107筆的資料未標註客戶的“性別Gender”。因此，我們無法確定是否已將更多產品出售給男性或女性。見答案D。
	D	Correct:	數據集的結構顯示有3 factors/levels: “”、“F”、“M”。這表明男性、女性和無指定或丟失。M顯示較高的值9,347，然後是F，8,872，但是，有45,107個損壞或丟失的數據，因此，我們不知道購買保險的男人數量是否比女人多。(附加資訊)
	E	Incorrect:	無法從圖3中確定，“Cancellation Plan”是最暢銷的產品。(附加資訊)
6	A	Incorrect:	這是年齡中位數(最常見)。中位數年齡顯示為36。(其他信息)
	B	Correct:	數據框顯示“Airlines” :17,457 和 “Travel Agencies” :45,869，“Travel Agencies”銷售的產品數量是“Airlines”的兩倍，這是正確的。(附加資訊)
	C	Incorrect:	第一個四分位數範圍Q1為35。第三個四分位範圍Q3為43，Q3 減去 Q1，IQR為8。(附加資訊)
	D	Incorrect:	這是不正確的。已經提出 927 項索賠，佔總銷售額 63,326，的 1.46%。(附加資訊)
	E	Correct:	“Cancellation Plan”的銷售量為 18,630，其次是“2 way Comprehensive Plan”，銷售量為13,158。(附加資訊)

Question: 1, Syllabus: IP, Part: D, Type: AR, SyllabusRef: IP0404, Level: 4			
1	False:	select() 函數允許您指定列。Filter 函數查看行。我們可以看到僅顯示了有限數量的列。同樣，“claim”列與此分析無關。(參考第54頁)	False: 使用select() 函數，可以根據您指定的任何條件選擇列。filter 函數幾乎與 select 函數完全相似，但它查看的是行(觀察值)而不是列。(參考第55頁)
2	True:	使用 arrange() 函數，您可以根據特定條件，對行進行重新排序。我們可以看到“Commission.in.value”已經排序了。(參考第55)	False: 使用re-name() 函數，您可以更改列名以保持一致，而無需對數據進行任何重大更改。(參考第56頁)
3	False:	如果NA數量有限，並且數據集足夠大，以至於減小之後，大小對總體結果影響不大，則此方法最合適。Drop() 函數將刪除63,326 行中的45,107。(參考第62頁)	True: drop() 函數將消除任何缺少值的行。(參考第62頁)
4	True:	如果數據集包含許多相似的觀察值，則fill() 函數特別適用。這種方法的好處是它不會使數據失真太多(與將NA替換為0相比)，因為它可以“平滑”觀察結果。但是，這種方法的缺點是結果完全取決於行的排序方式。不同排序的數據集將提供完全不同的結果。(參考第62頁)	True: fill()函數用缺失值上方的觀察值(行)來替換缺失值。但是，這種方法的缺點是結果完全取決於行的排序方式。不同排序的數據集將提供完全不同的結果。(參考第62頁) 這就是申論語句為真的原因，答案是A。

Question: 2, Syllabus: ER, Part: A, Type: MR, SyllabusRef: ER0301, Level: 3			
1	A	Incorrect:	對原始數據執行數據檢查操作，目的是清理數據並將其整理為所需要的格式。數據檢查操作在企業大數據管道 (Enterprise Big Data Pipeline) 的“準備”階段執行。在此階段，而不是在進行探索性數據分析，您將在其中積極致力於更改基礎數據集。(請參閱第36頁和第67頁)
	B	Correct:	探索型數據分析 (EDA) 的主要目標是了解數據中的變量，以便您可以看到一些模式 (some patterns, and form some initial ideas)，並就如何構建模型和演算法形成一些初步的想法。(請參閱第67頁)
	C	Incorrect:	對原始數據執行了數據檢查操作，目的是清理數據並將其整理為所需的格式。數據檢查操作在企業大數據管道(Enterprise Big Data Pipeline)的“準備”階段執行。在此階段，您正在積極地更改基礎數據集。(請參閱第67頁)
	D	Incorrect:	對原始數據執行了數據檢查操作，目的是清理數據並將其整理為所需的格式。數據檢查操作在企業大數據管道(Enterprise Big Data Pipeline)的“準備”階段執行。在此階段，您正在積極地更改基礎數據集。(請參閱第67頁)
	E	Correct:	探索性數據分析操作是在清理後的數據上執行的，目的是為了建立、完善或改進演算法和模型，並在企業大數據分析管道(Enterprise Big Data Pipeline)的“分析”階段執行。在此階段，您不再更新基礎數據集。(第67頁)
2	A	Correct:	直方圖可以將數據迅速分成幾個值範圍 (通常稱為“存儲桶buckets”)，以便您可以輕鬆查看數據集的分佈。從圖4的直方圖中，我們可以看到，女性購買保險的最常見年齡是35-40歲。(參考第68頁，附加資訊)
	B	Correct:	我們可以看到120歲左右存在一些離群值。(參考第68頁，附加資訊)
	C	Incorrect:	從圖4的直方圖中，我們可以看到購買保險的女性最普遍的年齡是35-40歲。 (“Age”的範圍是0到120，但是最年輕的女性顧客年齡在15到20歲之間。我們無法從該圖中確定年齡的中位數。參考第68頁，附加資訊)
	D	Incorrect:	圖4中的直方圖顯示，年齡在35至40歲之間的女性購買的保險產品最多。這並不能告訴我們該群體購買了哪些保險產品。(附加資訊)
	E	Incorrect:	從圖4的直方圖中，我們可以看到購買保險的女性最普遍的年齡是35至40歲。 (“Age”的範圍是0到120，但是最年輕的女性顧客年齡在15到20歲之間。我們無法從該圖中確定平均年齡。請參閱第68頁，附加資訊)
3	A	Incorrect:	圖中沒有顯示銷售產品的數量，只顯示了所有銷售產品的男女購買年齡分佈。(附加資訊)
	B	Correct:	Boxplots 也是一次非常有用的工具，它可以一次顯示多個變量，這有利於進行比較。從圖 5 的圖中，我們可以輕鬆確定，平均而言，購買保險產品的男性年齡比女性年齡大。您還可以看到四分位間距 (IQR) 的範圍更廣。(參考頁 669，附加資訊)
	C	Incorrect:	在圖 5 中，垂直線是中位數，而不是平均年齡。該圖顯示，女性的中位年齡在40歲以下，男性的中位年齡在40歲以上。(附加資訊)
	D	Correct:	圓圈表示女性離群點在80歲以上，男性離群點在85歲以上。還顯示了120歲的異常值。(請參見第69頁，附加資訊)
	E	Incorrect:	針對 30至50歲的女性銷售，圖上顯示的不是 75%，而是上下四分位數的範圍。
4	A	Incorrect:	知道變量是否趨於朝同一方向移動 (正相關或負相關) 會很有趣。散點圖是實現此目標的簡便可視化工具。在圖5中，我們不能真正說出“Age”和“Net.Sales”數據之間存在任何類型的關係。(請參見第71頁，附加資訊)
	B	Incorrect:	在圖5中，我們不能真正說出“Age”和“Net.Sales”數據之間存在任何類型的關係。(請參見第71頁，附加資訊)
	C	Incorrect:	無法利用散點圖確定IQR
	D	Correct:	散點圖非常有用，因為它們可以輕鬆地將兩個(或多個)變量相互比較。使用散點圖，您可以開始就數據集中的變量以及它們是否相關聯，建立一些初步的想法。圖5顯示了Y軸上低於“50”值就是大多數的數據點。(請參見第71頁，附加資訊)
	E	Correct:	y 軸上“250”的值上方顯示了 8 或 9 個數據點。(附加資訊)

Question: 2, Syllabus: ER, Part: B, Type: CL, SyllabusRef: ER0302, Level: 3			
1	A	Incorrect:	請參閱答案B.
	B	Correct:	目的是確定是否有足夠的證據可以拒絕原假設。通過拒絕原假設，它更有可能接受替代假設，這使替代假設可行。請注意，我們並不是說替代假設已被證明，而是原假設被拒絕了。(請參閱第73頁)
	C	Incorrect:	請參閱答案B
	D	Incorrect:	請參閱答案B
2	A	Correct:	觀察到的差異統計是實際值(觀察值)之間的差異，而不是任何預測值。在這種情況下，男性和女性的淨銷售額之間的差異。女性的數值("F"=59.2)減去男性的數值("M"=55.9)，=3.3。(第74頁，附加資訊)
	B	Incorrect:	請參閱答案A
	C	Incorrect:	請參閱答案A
	D	Incorrect:	請參閱答案A
3	A	Correct:	在隨機檢驗中，將觀察到的差異統計量與計算出的排列組合的差異統計量進行比較，看觀察到的差異統計量是否僅由偶然因素引起。(參考第74頁)
	B	Incorrect:	在隨機化測試的第2步中，計算出的差異統計量是針對每一個排列組合產生的。觀察到的差異統計量來自於原始數據集。(參考第74頁)
	C	Incorrect:	在原假設為真與變量獨立的假設下，用infer 套件產生排列。(參考第77頁)
	D	Incorrect:	請參閱答案A
4	A	Incorrect:	觀察到的差異統計量大於計算的差異統計量，因此可以拒絕原假設。(參考第79頁)
	B	Incorrect:	假設檢定並不能證明替代假設，否定原假設就更有可能接受替代假設。(參考第73頁)
	C	Correct:	從歷史以及行業標準的角度來看，拒絕原假設的閾值已設置為第5個百分點或第95個百分點。數據的“隨機”排列的95%應該低於此閾值。如果觀察到的差異統計量大於該值，則可以安全地拒絕該假設，因為該結果僅憑偶然發生的機會就很小。(參考第79頁)
	D	Incorrect:	請參閱答案C
5	A	Incorrect:	直方圖是直觀地理解此數據的典型方法，它應顯示第95個百分位數和觀察到的差異統計量。(參考第78頁)
	B	Incorrect:	這是不正確的。每個排列都沒有觀察到的統計差異。另請參閱答案C(there is not an observed statistical difference for each permutation)。(參見第74頁)
	C	Correct:	理解排列結果的最簡單方法之一是在圖形中查看它。通常，利用直方圖顯示帶有觀察到的差異統計量以及第5個和第95個百分位數的繪製的排列，是實現此目的的方法。如果觀察到的差異統計值高於第95個百分點或低於第5個百分點，我們可以拒絕原假設。(請參閱第78和79頁)
	D	Incorrect:	理解排列結果的最簡單方法之一是在圖形中查看它。通常，利用直方圖顯示帶有觀察到的差異統計量以及第5個和第95個百分位數的繪製的排列，是實現此目的的方法。如果觀察到的差異統計值高於第95個百分點或低於第5個百分點，我們可以拒絕原假設。(請參閱第78和79頁)
6	A	Incorrect:	請參閱答案C
	B	Incorrect:	請參閱答案C
	C	Correct:	統計推論的理論和技術已廣泛用於企業大數據分析中，因為統計推論的目的是根據特定樣本估算未來的不確定值。(參考第72)
	D	Incorrect:	請參閱答案C

Question: 2, Syllabus: ER, Part: C, Type: CL, SyllabusRef: ER0403, Level: 4			
1	A	Correct:	利用散點圖，您可以開始形成有關數據集中變量的一些初步構想。如果圖形上的直線從左到右向上移動，則它具有正線性關係。如果圖形上的直線從左到右向下移動，則它具有負線性關係。在此圖中，數據中沒有顯示向上或向下趨勢的模式。(請參閱第71和82頁)
	B	Incorrect:	關聯很有用，因為它們可以指示出在實際情況下可以利用的變量之間，預測關係。利用散點圖的檢查，可以確定變量之間是否可能存在線性關係。(請參閱第83頁)
	C	Incorrect:	請參閱答案 A
	D	Incorrect:	請參閱答案 A
2	A	Incorrect:	請參閱答案 B
	B	Incorrect:	關聯係數(correlation coefficient) +1 表示完美的正線性關係，關聯係數 -1 表示完美的負線性關係。如果線變為正無窮大，則y預測將變為1。如果線變為負無窮大，則y預測將變為0。(請參閱第83頁)
	C	Correct:	正相關是兩個變量之間的關係，以使它們的值一起增加或減少。在一個完美的正相關中(表示為+1)，一個變量的增加或減少總是預測第二個變量的方向變化相同。(請參閱第83頁)
	D	Incorrect:	負相關描述了兩個變量沿相反方向移動的程度。例如，對於X和Y這兩個變量，X的增加與Y的減少相關。在此題例中，實際值表示正相關，即“Net.Sales”(X)傾向於增加帶有“Duration”(Y)。(請參閱第83頁)
3	A	Incorrect:	請參閱答案 B
	B	Correct:	關聯性是有用的，因為它們可以表明變量之間的預測關係，可以在實際情況下加以利用。然而，關聯性的存在並不足以推斷出因果關係的存在。兩個變量之間存在統計關係的事實並不一定意味著一個變量是由另一個變量引起的。(請參閱第83頁)
	C	Incorrect:	請參閱答案 B
	D	Incorrect:	請參閱答案 B
4	A	Incorrect:	值0.3143確實顯示弱的正相關。正相關是兩個變量之間的關係，以使它們的值一起增加或減少。(請參閱第83頁)
	B	Incorrect:	值為0.3143確實顯示弱的正相關。正相關是兩個變量之間的關係，以使它們的值一起增加或減少。負相關描述兩個變量在相反方向上移動的程度。(請參閱第83頁)
	C	Incorrect:	關聯性是有用的，因為它們可以表明變量之間的預測關係，可以在實際情況下加以利用。然而，關聯性的存在並不足以推斷出因果關係的存在。兩個變量之間存在統計關係的事實並不一定意味著一個變量是由另一個變量引起的。(請參閱第83頁)
	D	Correct:	儘管微弱，但正相關是兩個變量之間的關係，以使它們的值一起增加或減少。在一個完美的正相關中(表示為+1)，一個變量的增加或減少總是預測第二個變量的方向變化相同。(請參閱第83頁)

Question: 2, Syllabus: ER, Part: D, Type: AR, SyllabusRef: ER0404, Level: 4

1	True:	在簡單的線性回歸中，單一自變量用於預測因變量的值。(參考第84)	True:	在簡單線性回歸中，一個單一的自變量被用來預測一個因變量的值。因此，簡單線性回歸的目的是預測性的，並且經常在其中一個變量(自變量)可以被影響的時候使用。(參考第84頁) 這就是申論語句為真的原因，答案是A。
2	True:	在簡單線性回歸中，一個單一的自變量被用來預測一個因變量的值。因此，簡單線性回歸的目的是預測性的，並且經常在其中一個變量(自變量)可以被影響的時候使用。(參考第84頁)	False:	在簡單線性回歸中，一個單一的自變量被用來預測一個因變量的值。因此，簡單線性回歸的目的是預測性的，並且經常在其中一個變量(自變量)可以被影響的時候使用。(參考第84頁)
3	True:	正相關描述了兩個變量之間的關係，即它們的值一起增加或減少。負相關描述的是兩個變量向相反方向移動的程度。(參考第83頁)	True:	關聯性是確定兩個變量的共同關係或關聯的一種統計度量。回歸描述了一個自變量與因變量在數值上的關係。係數是否為負，不關乎關聯性的強弱，而關乎一個變量是否隨著另一個變量的減少而增加，所以答案為B，(參考第83頁)
4	False:	最佳擬合線是通過散點圖上的最大數量的點畫出的一條直線，平衡線上和線下大約相同數量的點。(參考第85頁)	True:	回歸線有時被稱為“最佳擬合線”，因為它是在畫過各點時最適合的線。它是一條使平方殘差(因變量的實際值和預測值之間的差異)之和最小的線(參考第85頁)
5	True:	線性關係可以利用給定值X，來預測Y的(平均值)數值，使用一條直線(稱為回歸線)。如果你知道這條回歸線的斜率和y-截距，那麼你就可以插入一個X的值，並預測Y的平均值 (參考第87頁)	False:	“最佳擬合”是指利用點，畫出的最適合的線。它是一條使平方殘差(因變量的實際值和預測值之間的差異)之和最小的線。該線與y軸交叉的點就是y截距 (參考第85頁)
6	True:	無論你在數據中應用的演算法多麼複雜，它永遠是現實的近似值(。可以將數學公式或模型應用於數據，以確定變量之間的關係，比如可以根據數據中的其他變量，開發出關聯或回歸模型來評價數據中的某個變量，根據模型的準確性，會有一些殘餘誤差。(參考第20頁和第66頁)	True:	平方殘差的總和是因變量的實際值和預測值之間的差異(。它用於找到“最適合”的線性回歸線。這與殘差誤差值不同，殘餘誤差值是一個模型準確性的指標。(參考第85頁)。故選擇答案是 B。

Question: 3, Syllabus: CM, Part: A, Type: CL, SyllabusRef: CM0301, Level: 3

1	A	Incorrect:	請參閱答案 B
	B	Correct:	k-NN分類器總是需要有兩個數據集。訓練數據集用於學習，有助於確定應該為每個新的觀測值分配到哪些類別。測試數據集用於測試分類器是否真的有效。有了測試數據集，我們可以使用比較k-NN演算法的預測值和原始測試數據的標籤來確定模型的準確性。在附加資訊中，測試數據集顯示為觀測值60001:63326，測試集中有3326個觀測值，將對其進行分類預測。(參考第94頁)
	C	Incorrect:	請參閱答案 B
	D	Incorrect:	請參閱答案 B
2	A	Incorrect:	請參閱答案 D
	B	Incorrect:	請參閱答案 D
	C	Incorrect:	請參閱答案 D
	D	Correct:	在計算預測值時，沒有指定k的值，所以默認值將會被使用。k-NN分類器的默認值 $k = 1$ 。(參考第94頁)
3	A	Incorrect:	k-NN 分類器的工作原理是它計算數據點之間的歐氏距離。為了平等地對待所有變量，進行這種分類練習的最好方法是將所有變量的尺度重新調整到 0-1 的範圍，這個過程被稱為常態化。所有的值(在訓練集以及測試集中)，都需要進行常態化。(參考第90頁)
	B	Incorrect:	隨著數據集越來越大(尤其是增加了更多“品質好”的標籤數據時)，模型會越來越準確，並學會如何做出更準確的決策。(參考第97頁)
	C	Correct:	選擇的結果，以及一個點是否被正確分類，都取決於k的值，那麼k的好值是多少呢？不幸的是，這個問題沒有統一或正確的答案，因為它取決於數據集的結構、變量的數量和分佈。最好的方法之一是用多個版本的 k 來運行實驗，並確定哪個 k 的精度最高(即數據的數量點分類正確)(參考第94頁)
	D	Incorrect:	請參閱答案 C
4	A	Incorrect:	請參閱答案 D
	B	Incorrect:	請參閱答案 D
	C	Incorrect:	請參閱答案 D
	D	Correct:	因為 k-NN 的工作原理是基於計算數據點之間的距離，所以有一個因素始終需要被考慮。不同變量的取值範圍會對最終結果產生重大影響。為了平等地對待所有變量，將所有變量重新縮放到 0-1 的範圍，這個過程被稱為常態化(標準化)。所有的值(無論是訓練集還是測試集)，都需要進行常態化。(參考第96頁)
5	A	Incorrect:	請參閱答案 D
	B	Incorrect:	請參閱答案 D
	C	Correct:	當我們比較不同的 k 值的不同準確度水平時，我們可以看到，我們的模型的準確度不斷提高，直到 $k=7$ ，之後又開始下降。根據這一觀察，該分類模型中 k 的最佳值是 $k=7$ 。(參考第96頁)
	D	Incorrect:	請參閱答案 C

Question: 3, Syllabus: CM, Part: B, Type: CL, SyllabusRef: CM0302, Level: 3

1	A	Incorrect:	與k-NN分類器可用於距離計算不同，單純貝葉斯可用於計算概率。它計算出未知變量屬於某個類別的最大可能性。因此，它不需要常態化的數據。(參考第98頁)
	B	Incorrect:	單純貝葉斯分類器的主要優點之一是它能夠快速計算多個變量的分類預測(參考第105頁)
	C	Correct:	單純貝葉斯演算法是一種計算友好型演算法，能夠快速進行計算和預測。因此，單純貝葉斯分類器經常用於進行即時預測，尤其是在智慧手機應用中。(參考第98頁)
	D	Incorrect:	單純貝葉斯模型是一種用於解決預測業務目標的預測分類技術(參考第27頁)
2	A	Incorrect:	混淆矩陣是以預測類與實際類的交叉表的形式產生成的。它是使用predict()和table()函數所產生的，並且與概率表分開產生。(參考第105和121頁)
	B	Incorrect:	可以在維恩圖中顯示聯合概率(而非條件概率)，但這不是概率表的一部分。(參考第100頁)
	C	Correct:	概述是概率表的一部分。這顯示了購買每種GGI產品的先驗概率，但是缺少“Gender”和“Product”以及“Age”和“Product”的聯合概率的條件概率。(參考第103、104頁，附加資訊)
	D	Incorrect:	概率表未顯示實際分類與預測分類。這是使用predict()和table()函數分別產生的。(參考第103頁)
3	A	Incorrect:	請參閱答案B
	B	Correct:	如果值之一為零，則在相乘時，所有其他變量也將為零。當然，這不是對現實的充分反映，因為可能仍會發生這種類型變量組合(可能很小)的情況。為了克服單純貝葉斯演算法的這一問題，我們可以為每項結果組合添加一個虛構數(通常為1)。我們稱此解決方案為拉普拉斯平滑。拉普拉斯平滑處理可確保始終存在不同變量之間的很小重疊，因此無法完全排除任何結果。(參考第107頁)
	C	Incorrect:	如果數據集不包含此組合的例子，則條件概率將顯示為零。拉普拉斯平滑可用於克服此問題(參考第107頁)
	D	Incorrect:	將預測添加到原始數據集中，因此可以輕鬆比較實際類別和預測類別，以查看模型的準確性。(參考第105頁)
4	A	Incorrect:	組合人口統計和地理變量不會降低模型的預測能力。單純貝葉斯假設變量是獨立的，並且能夠計算多個變量的類別預測。(參考第105頁)
	B	Incorrect:	單純貝葉斯假設變量是獨立的，並且能夠計算多個變量的類別預測。(參考第105頁)
	C	Correct:	單純貝葉斯假設變量是獨立的，並且能夠計算多個變量的類別預測。添加變量可以增強模型的預測能力。(請參閱第105和106頁)
	D	Incorrect:	單純貝葉斯假設變量是獨立的，並且能夠計算多個變量的類別預測。(參考第105頁)
5	A	Incorrect:	將Laplace參數添加到naïve-bayes()函數會處理空值(如果將其設置為=1)。這與計算模型的準確性無關。值為0表示沒有拉普拉斯平滑。(參考第107頁)
	B	Incorrect:	將條件概率與先驗概率進行比較不會提供有關模型準確性的任何資訊。為了評估這一點，對整個數據集使用了predict()函數來比較實際類別和預測類別。(參考第104頁)
	C	Incorrect:	重複執行該函數不會影響朴素貝葉斯模型的準確性。另請參閱答案D
	D	Correct:	為了進行比較，您可以將預測添加到原始數據集中，這樣就可以輕鬆比較實際類別和預測類別。新列將添加到您的數據集中，以指示預測的類別。如果隨後將預測類別與實際類別列表在一起，則可以使用均值運算來檢查模型的準確性。(參考第105頁)

Question: 3, Syllabus: CM, Part: C, Type: AR, SyllabusRef: CM0403, Level: 4

1	False:	在邏輯回歸中，我們可以指定哪些可能具有預測作用的變量，我們希望將其納入模型中。這些輸入變量在邏輯回歸模型中可以是分類、整數或數字變量，目標變量必須有一個二進制值--例如可以用1或0，或Yes或No來表示 (請參閱第110頁)	True:	邏輯回歸中目的是基於邏輯函數尋找參數。在邏輯模型中，標記為“1”的值的對數奇數(賠率的對數)是一個或多個自變量(“預測變量”)的線性組合。變量的聯合概率以0到1之間的分數形式返回。(請參閱第111頁)
2	True:	此數據感興趣的二進制變量是“claim”變量，它指示客戶是否已經提出索賠。(用1表示) 或沒有(用0表示)。在函數“glm()”之後首先指定它。(參考第110頁)	True:	邏輯回歸模型的正確輸入變量可以是分類變量、整數變量或數字變量，但這不能解釋為什麼該模型中感興趣的二進制變量是 Claim_Data。因此，答案是B
3	True:	有必要將 predict() 函數應用於模型對象以預測未來的行為。使用predict() 函數，您可以在數據中添加另一列，以估計客戶提出索賠的可能性。(參考第113頁)	True:	該模型本身只是根據輸入對輸出的概率進行建模，並且不執行統計分類(它不是分類器)，雖然它可以用來做一個分類器，選擇一個截止值，把概率大於截止值的輸入歸為一類，低於截止值的歸為另一類。(請參閱第115頁) 這就是申論語句為真的原因，答案是A。
4	True:	我們知道，有1.5%的GGI客戶實際提出過索賠。使用最喜歡產生預測影響的變量，邏輯回歸模型計算這些客戶提出索賠的可能性。將分類閾值設置為最高1.5%的概率，即0.2189，可以使用此模型來識別可能提出索賠的未來客戶(參考第115頁)	True:	知道樣本集中實際分類的概率得分，將影響您設置的對未來預測進行分類的閾值(請參閱第115頁)。這就是申論語句為真的原因，答案是A。
5	False:	1.5%是在原始數據集中提出索賠的人數 - 該數字可用於設置概率閾值 - 不是模型的準確性數字。這是利用預測結果與數據集中的實際結果進行比較來確定的。將“閾值”設置為高於或低於觀察到的統計資訊將影響預測的分類數量。儘管不可能建立一個完美的模型，但將閾值設置為與實際索賠人的可能性越接近，則可能越準確。在這種情況下，分析人員已將實際數據中顯示的百分比用作預測模型的閾值，因此準確度應盡可能接近100%(參考第115頁)	True:	將閾值設置得太高只會將一個很小的百分比表示為“True”(可能存在更多真實的風險)，而將閾值設置得太低會表明太多的結果表示為“True”，也就是說要麼太少，否則太多預測案例將被錯誤分類。(請參閱第115頁)。
6	False:	要確定邏輯回歸模型的準確性，您需要將預測列與保險產品數據集中的實際索賠數據進行比較(使用table()函數)(To。閾值集決定了多少百分比應被視為預測的類別。雖然閾值集會對模型的準確性產生影響，但它並不能決定它。(參考第115頁)	False:	通過predict()函數，某人提出索賠的可能性被指定為一個百分比，這取決於您確定的閾值，這個閾值表示百分比應該是多少，才會認為某人是索賠者或不是索賠者樣本集的實際分類百分比與閾值集的百分比之間的差異，示將被錯誤分類的預測數量，它不是模型的準確性值。模型的準確性是通過比較預測結果和數據集中的實際結果來確定的(參考第115頁)

Question: 3, Syllabus: CM, Part: D, Type: CL, SyllabusRef: CM0403, Level: 4

1	A	Incorrect:	請參閱答案 B
	B	Correct:	分類樹演算法首先在變量上對數據集進行分割，這將導致最相同性質的子組，其中決策節點將盡可能多的相似觀察進行分組。該標準不是分析師輸入的 <code>rpart()</code> 函數的參數。(參考：第117頁)
	C	Incorrect:	請參閱答案 B
	D	Incorrect:	請參閱答案 B
2	A	Incorrect:	操作是適當的。 <code>maxdepth</code> 被指定為函數中的參數。它指定演算法何時停止以及最終樹節點的最大深度。(請參見第120和123頁)
	B	Incorrect:	操作適當。過度擬合是指學習系統與給定的訓練數據緊密配合，以至於在預測未訓練數據的結果時不準確的現象。在分類樹中，例如當未指定最大深度值，同時樹被設計為完全適合訓練數據集中的所有樣本時，就會發生過度擬合 (參考第121頁)
	C	Correct:	預修剪從一開始就確定了樹的最大尺寸可能會變成多少。例如，您可以確定決策節點的最大數量，或者確定應該作為葉節點一部分的觀察值的最小數量。可以使用 <code>maxdepth</code> 指定預修剪， <code>maxdepth</code> 設定任何節點的最大深度。(參考第123頁)
	D	Incorrect:	操作正確，請參閱答案 C。但是數據集中的變量數量並不能確定 <code>maxdepth-set</code> 。(參考第123頁)
3	A	Incorrect:	<code>Cp</code> 圖顯示，用於最佳精度的良好複雜度參數約為0.0047，其中，複雜度在虛線以下。儘管這大約是 17棵樹的大小，但是數據集 (15)中變量的數量與是否使用 <code>Cp</code> 參數無關。(參考第124頁)
	B	Incorrect:	<code>Cp</code> 圖顯示了用於獲得最佳精度的複雜參數。 <code>Cp</code> 圖中圖的形狀表明，樹的最佳大小約為0.0047，即複雜度在虛線以下。(參考第124頁)
	C	Incorrect:	<code>Cp</code> 圖表明了對於不同大小的樹之模型的準確性。它是從完全生長的樹產生的，指示可應用於修剪的複雜性參數。在這種情況下，最佳精度約為0.0047，樹的大小約為 17 (參考第124頁)
	D	Correct:	<code>Cp</code> 圖顯示了用於獲得最佳精度的複雜參數。 <code>Cp</code> 圖中圖的形狀表明，樹的最佳大小約為0.0047，即複雜度在虛線以下。(參考第124頁)
4	A	Correct:	後修剪讓樹首先生長到完整大小，然後通過指定 <code>Cp</code> 複雜性參數將其修剪回所需大小。這將複雜度變量設置為任何其他數值 (因此是一個更複雜的模型) 都不會顯著提高模型準確性的程度。您可以通過使用圖形複雜性參數操作來繪製模型的複雜度來確定適當的 <code>cp</code> 點。在此 <code>Cp</code> 圖中，最佳 <code>Cp</code> 值應為0.0095，最大深度約為15。(請參閱第124頁)
	B	Incorrect:	<code>Cp</code> 圖顯示了用於獲得最佳精度的複雜參數，它沒有測量實際精度。這是用混淆矩陣來衡量的。(參考第124頁)
	C	Incorrect:	模型應該在不同的數據集上進行測試，即測試集與訓練集相比的結果，準確率達到 82.9%的數字來自 <code>maxdepth</code> 參數為10的第一個模型。然後，可以使用 <code>Cp</code> 圖來查看是否可以通過使用複雜參數產生模型來提高精度。(參考第124頁)
	D	Incorrect:	82.9% 的準確度數據來自第一個使用 <code>maxdepth</code> 參數10修剪的模型。但是，可以使用 <code>Cp</code> 圖來查看是否可以使用複雜參數產生模型來提高準確性。(參考第124頁)

Question: 4, Syllabus: CV, Part: A, Type: MR, SyllabusRef: CV0301, Level: 3

1	A	Correct:	與分類(在上一節中討論過)不同，聚類是非監督學習的一個方法。沒有首先送入機器的樣本數據，但是電腦會根據組之間的相似性開始製定聚類。(參考第125頁)
	B	Incorrect:	層次聚類模型適用於數字、二進制或有序數據，但是使用了不同的技術來測量數據之間的相似性或不相似性。歐幾里得距離主要適用於數字數據，例如產品的價格，溫度變化等。當數據為二進制或有序數據時，我們可以使用 Jaccard 指數。(參考第126頁)
	C	Incorrect:	與分類(在上一節中討論過)不同，聚類是非監督學習的一個方法。沒有首先送入機器的樣本數據，但是電腦會根據組之間的相似性開始製定聚類。(參考第125頁)
	D	Incorrect:	有關大數據集的第一個問題是要考慮數據中的缺失值(missing values)。由於層次聚類使用 distance() 函數來計算數據點之間的距離，因此丟失數據點 (NA) 將導致錯誤(參考第136頁)
	E	Correct:	可以將分層聚類應用於具有多個變量的大數據集。但是，重要的是要注意，每個變量都可能對結果產生嚴重影響。重要的是要處理異常值、數據丟失以解決維度差異。(參考第114、115頁)
2	A	Incorrect:	請參閱答案 B
	B	Correct:	層次聚類演算法使用測量數據觀測值之間的距離來工作。該演算法首先選擇最靠近的兩個點，然後迭代地將數據點添加到與前一組具有最接近距離的數據點。(參考第126頁)
	C	Incorrect:	請參閱答案 B
	D	Incorrect:	在層次聚類中，根據數據點的歐幾里得距離或 Jaccard 指數將數據點分組在一起的過程一直進行到所有點都被分組為止。(參考第128頁)
	E	Correct:	層次聚類演算法使用測量數據觀測值之間的距離來工作。(參考第126頁)
3	A	Incorrect:	在k均值聚類中使用 Elbow 方法來確定k的值是最優的值。在層次聚類中，k表示最終結果中需要指定的聚類數。聚類數可以預先確定(使用k)，也可以根據聚類之間的距離(使用高度)在事後確定。(參考第127,145頁)
	B	Correct:	因為數據需要標準化，所以將採用數值形式(請參閱答案D)，因此可以使用歐幾里得距離來構建模型。(參考第127、136頁)
	C	Incorrect:	在此案例中，可以對“Age”、“Duration”與“Spend”進行標準化，並為其分配數值，從而可以使用歐幾里得。如果沒有數值可分配給數據點，則使用 Jaccard 指數，該統計用於衡量樣本集的相似性和多樣性。(參考第136頁)
	D	Correct:	由於我們使用 distance() 函數來測量數據點之間的距離，因此變量的尺寸非常重要。不考慮尺寸可能會導致結果偏斜或偏差，以及不準確的聚類。為了確保所有數據都是無單位的，我們可以將標準化過程應用於所有變量。在此數據集中，保險產品的支出會因 age 與 duration 而有所不同維度。(參考第136頁)
	E	Incorrect:	推論是對母體參數和統計關係的可靠性進行判斷的過程，通常是基於隨機抽樣(參考第72頁)
4	A	Correct:	選擇不同的鏈接標準可能會導致模型產生不同的結果。鏈接方法指定以哪種方式計算群集之間的距離。完整鏈接：兩個群集之間的最大距離，平均鏈接：兩個群集之間的平均距離。平均鏈接方法不太可能受任何異常值的影響，但因此可能無法檢測到細微的模式。(參考第130頁)
	B	Incorrect:	平均鏈接的方法不太可能受任何異常值的影響，但是可能無法檢測到細微的模式。(參考第130頁)
	C	Correct:	鏈接方法指定以哪種方式計算群集之間的距離。完整鏈接：兩個群集之間的最大距離，平均鏈接：兩個群集之間的平均距離。平均鏈接方法不太可能受任何異常值的影響，但因此可能無法檢測到細微的模式。(參考第130頁)
	D	Incorrect:	完整鏈接法和平均鏈接法適用於數值。(請參閱第133和137頁上的案例)
	E	Incorrect:	鏈接方法指定以哪種方式計算群集之間的距離。完整鏈接：兩個群集之間的最大距

			離，平均鏈接：兩個群集之間的平均距離。平均鏈接方法不太可能受任何異常值的影響，但因此可能無法檢測到細微的模式。(參考第130頁)
5	A	Incorrect:	在進行預修剪時會在“分類”樹中使用設置最大深度，以儘早停止樹的生長過程。設置最終樹的任何節點的最大深度，並將根節點計為深度0。(請參閱第123頁)
	B	Correct:	群集數 (k)：在最終結果中，所需要指定的群集數。群集數數可以預先指定(使用k)，也可以根據群集之間的距離(使用高度)，在事後確定。(參考第127頁)
	C	Correct:	可以預先指定群集數(使用k)，也可以根據群集之間的距離(使用高度)在事後確定。(參考第127頁)
	D	Incorrect:	n不是用於層次結構群集的參數。請參閱答案 B.
	E	Incorrect:	分類樹：預修剪會儘早停止樹的生長過程，並且可以使用 Minsplit 設置節點中必須存在的最小觀察數，才能嘗試進行拆分。此方法禁止在決策節點的觀察數少於指定數目時拆分決策節點。(參考第123頁)
6	A	Incorrect:	平均值較低的群集不是引起關注的原因，儘管行銷組可能會對該組的內容感興趣，例如一個年齡較小，持續時間短且花費少的群體可能具有這種特徵。行銷部門可能希望針對該群體的產品提供與另一集群中的客戶不同的產品。(參考第137頁)
	B	Correct:	小聚類中的極端值表示異常。這些應該進行審查和進一步檢查(參考第137頁)
	C	Incorrect:	標準化的結果是，每個變量的平均值為0，標準偏差為1。使用normalize() 函數時，可以在匯總統計資訊中看到這一點：它不是匯總表的一部分最終數據。(參考第137頁)
	D	Incorrect:	標準化的結果是，每個變量的平均值為0，標準偏差為1。使用normalize() 函數時，可以在匯總統計資訊中看到這一點：它不是匯總表的一部分最終數據。(參考第137頁)
	E	Correct:	小聚類中的極端值表示異常。這些應該進行審查和進一步檢查。(參考第137頁)

Question: 4, Syllabus: CV, Part: B, Type: CL, SyllabusRef: CV0302, Level: 3			
1	A	Correct:	變量必須全部為數字，k-means演算法才能起作用。因為我們使用distance() 函數(原則1)來測量數據點之間的距離，所以變量的大小非常重要。不考慮尺度可能會導致結果偏斜或偏差，以及不正確的聚類。為了確保所有數據都是無單位的，我們可以將標準化過程應用於所有變量。(參考第141頁)
	B	Incorrect:	在k均值聚類中，每個數據點均基於到 k 質心的最短歐幾里得距離分配給一個聚類。(參考第140 頁)
	C	Incorrect:	數據經過標準化處理，以避免尺度出現偏斜或誤差，不是因為變量的均值必須為1。但是，由於標準化的結果，所有變量的均值都將為0。(請參閱第136和141頁)
	D	Incorrect:	請參閱答案 A
2	A	Incorrect:	如果未指定 k，則是層次聚類算法而不是k-means，它將繼續根據數據點之間的距離對聚類進行分組，直到將所有點分組為一個聚類為止。(參考第127頁)，請參閱答案 D。
	B	Incorrect:	這兩種演算法產生的聚類數取決於k的值。在層次聚類中，可以預先或之後指定k。在 k 均值聚類中，必須事先指定 k。(參考第127和142 頁)
	C	Correct:	當實例到集群的分配沒有進一步變化時，k-means演算法收斂(請參閱第140頁)，請參閱答案 D。
	D	Incorrect:	鏈接方法是分層聚類演算法使用，而不是 k-means 演算法使用。鏈接方法指定計算聚類之間距離的方式。(參考第127頁)
3	A	Correct:	k-均值聚類的目標是在將n個觀察值劃分為k個聚類，其中每個觀察都屬於具有平均值最近的聚類每個數據點都基於距 k 質心的最短歐幾里得距離，分配給一個聚類。質心已更新。質心的新位置是群集中所有數據點的中心。(參考第139頁)
	B	Incorrect:	每個數據點都基於距 k 質心的最短歐幾里得距離，分配給一個聚類質心已更新。質心的新位置是群集中所有數據點的中心(參考第140頁)
	C	Incorrect:	在層級聚類演算法中用於定義聚類數量的一種方法是指定高度，即聚類之間的最大距離，作為確定數據點聚類方式的分界線。(參考第127頁)
	D	Incorrect:	使用k值隨機產生“虛構”的質心。與相同性質子組有關的是分類樹演算法。(參考第117和140頁)
4	A	Incorrect:	k是手動設置的，它確定將數據點分配給的cluster數。使用k-means算法時，始終需要從頭開始選擇k。(參考第140和142 頁)，請參閱答案B。
	B	Correct:	使用彎頭法確定k的最佳值。彎頭方法的想法是在一個數據集中，針對k個不同的k值運行k均值聚類，並為k的每個值計算平方誤差和 (SSE)。然後，針對每個k值繪製SSE折線圖。如果折線圖看起來像一條手臂，那麼手臂上的“肘”就是最好的k值。此時，SSE開始趨於平坦，這意味著添加其他群集不會進一步增加群集之間的差異。(參考第144頁)
	C	Incorrect:	請參閱答案 B
	D	Incorrect:	層次聚類演算法具有樹狀結構，並且組合了最相似的聚類。這並不是由分析師所指定。k均值聚類算法是一種迭代結構，最相似的數據點被組合成為聚類，並且聚類之間盡可能的相似。(參考第144頁)
5	A	Incorrect:	請參閱答案 C
	B	Incorrect:	請參閱答案 C
	C	Correct:	理想的彎頭法是針對一系列不同的k值在數據集中運行k均值聚類，並針對k的每個值計算平方誤差總和 (SSE)。然後，針對每個 k 值繪製SSE折線圖。如果折線圖看起來像一條手臂，那麼手臂上的“肘”就是最好的 k 值。此時，SSE開始變得平坦，這意味著添加其他群集不會進一步增加群集之間的差異。(參考第144頁)
	D	Incorrect:	請參閱答案 C

Question: 4, Syllabus: CV, Part: C, Type: AR, SyllabusRef: CM0403, Level: 4

1	False:	在使用Grubbs檢驗時，數據分析師沒有預先指定單個數據點與其他數據之間的距離，然後再將其視為異常值。儘管對此的絕對答案取決於問題的背景和領域的專業知識，但我們可以使用Grubbs檢驗作為一個相當有力的指標來表明數據點是否是異常值。Grubbs檢驗適用於常態分佈。從均值超過三個標準差的任何數據點都可能表示數據不正確或有缺陷。(參考第149頁)	True:	Grubbs檢驗在數據隨機分佈且呈常態分佈的假設下工作。(參考第149頁)
2	True:	較小的p值(通常小於0.05)表示有力的證據反對原假設，因此您可以拒絕原假設。p值0.333遠大於此值，因此不太可能成為異常值。(參考第149頁)	True:	基於標準化常態分佈的屬性，Grubbs檢驗運行統計假設檢定，離群值(無論是最大值還是最小值)都應落入數據的常態分佈範圍內。因此，用Grubbs檢驗測量p值是正確的。較小的p值(通常 > 0.05)表示反對原假設的有力證據，因此您可以拒絕原假設。(參考第149頁) 原因不能解釋為什麼值0.333會使數據指向異常值，因此答案是B。
3	False:	Grubbs檢驗是一種用於檢測假設來自常態分佈母體的單變量數據集中的異常值的檢驗。Grubbs測試只能測試單個變量上是否存在異常值。(參考第125頁和第149頁)	False:	雖然在探索型分析中使用直方圖來檢查Grubbs檢驗是否可以有效地進行，但它們並不是實際檢驗的一部分。同樣，直方圖將數據分為多個值範圍(存儲桶)，以便您可以輕鬆查看數據集的分佈。它使用矩形以相等大小的連續數值間隔顯示數據項的頻率。它不同於條形圖，在某種意義上，條形圖涉及兩個或多個變量，而直方圖僅涉及一個變量。(參考第149頁)
4	False:	Grubbs 檢驗適用於常態分佈。從均值超過三個標準差的任何數據點都可能表示數據不正確或有缺陷。(參考第149頁) 它是k-NN離群值檢測演算法，可計算歐幾里得距離以檢測離群值。與其周圍鄰居的(平均)距離越高，數據點越可能是異常值。(參考第152頁)	True:	在散點圖中，可以使用k-NN離群值檢測演算法將看似獨立數據點的單個數據點識別為異常值。出現在非常寬的直方圖最右邊的單個數據點很可能是異常值。(請參閱第150和152頁)

Question: 4, Syllabus: CV, Part: D, Type: CL, SyllabusRef: CV0403, Level: 4

1	A	Incorrect:	Codebook 中記錄的資訊可能需要更改，例如有關數據的註釋或評論的更新，這對希望重現分析的人員將有所幫助。利用良好的版本控制，對Codebook的更改不會使它的有效性失效。如果人們可以使用該Codebook，或者可以更改其內容，則重要的是要保持良好的版本控制措施，以充分跟蹤誰進行了更改。(參考第157頁)
	B	Incorrect:	請參閱答案 D
	C	Incorrect:	Codebook 描述了數據集合的內容、結構和佈局，以作為分析中使用的數據記錄以及應如何解釋該記錄。它不會獲得完成分析所花費的時間。(參考第157頁)
	D	Correct:	記錄良好的 Codebook，包含的資訊，主要是在對於呈現數據檔案中的每個變量完整且易於解釋。它確保可以復制相同的分析。如果人們可以使用該 Codebook，或者可以更改其內容，則重要的是要保持良好的版本控制措施，以充分跟蹤誰進行了更改。(參考第157頁)
2	A	Correct:	缺失的數據可能會以“N/A”值或零的形式出現，因此區分兩者非常重要。應當記錄丟失數據的值和標籤。(參考第157頁)
	B	Incorrect:	雖然可以在探索型數據分析中以及開發某些演算法時完成，但仍不會從原始數據集中刪除丟失的數據。因此，重要的是應記錄丟失數據的值和標籤。(參考第157頁)
	C	Incorrect:	記錄每個變量的缺失值數量還不夠。丟失的數據可能以“N/A”值或零的形式出現，因此區分兩者非常重要。應當記錄丟失數據的值和標籤。(參考第157頁)
	D	Incorrect:	記錄每個變量的缺失值數量還不夠。丟失的數據可能以“N/A”值或零的形式出現，因此區分兩者非常重要。應當記錄丟失數據的值和標籤。(參考第157頁)
3	A	Correct:	可視化－像似數據集－的建立方式在可視化的含義上已達成共識。因為全世界的每個人都以相同的方式使用和解釋可視化，所以可視化變得非常有效。(參考第158頁)
	B	Incorrect:	請參閱答案 A
	C	Incorrect:	請參閱答案 A
	D	Incorrect:	請參閱答案 A
4	A	Incorrect:	請參閱答案 C
	B	Incorrect:	請參閱答案 C
	C	Correct:	與主要涵蓋1個變量的條形圖和直方圖不同，散點圖對於涵蓋至少兩個變量非常有用。在指定外觀時，可以通過使用顏色將另外一個變量添加為另一個維度。(參考第163頁)
	D	Incorrect:	散點圖無法比較多組數據點。這是集群的描述。(參考第163頁)
5	A	Incorrect:	無論幾何形狀還是美學外觀都發生變化，兩個可視化中的數據點都相同。但是，由幾何形狀決定用於顯示數據的形狀和視覺元素。(參考第158頁)
	B	Correct:	是幾何形狀決定了用於顯示數據的形狀和視覺元素。美學決定了繪製數據的畫布和軸。(參考第158頁)
	C	Incorrect:	雖然美學確實顯示了變量如何映射到繪圖的幾何圖形(畫布和軸)，但條形圖和直方圖卻具有不同的幾何圖形。需要改變的是幾何形狀而不是美學。(參考第158頁)
	D	Incorrect:	請參閱答案 D