参考答案

项目一　初识数据

一、选择题

1. B　　2. C　　3. A　　4. A

二、填空题

1. 符号记录　　信息　　知识
2. 定量数据　　数量数据
3. 确定目标　　划定边界
4. 原始数据
5. 推断性分析

三、简答题

1. 数据分析是指采用适当的方法和技术对收集得到的数据进行探索和分析，以提取有价值的信息并形成结论或知识的一系列过程。其目的是把隐藏在一大批看起来杂乱无章的数据中的信息萃取和提炼出来，以找出所研究对象的内在特性或规律。在实践中，数据分析可帮助人们定量地做出判断，以便采取适当的决策和行动。

2. 略

3. 问题定义是确保数据分析过程有效性的一个基础环节，主要包含两部分内容：确定目标和划定边界。

实际中的问题往往是由目标与现实之间存在的差距或矛盾所产生的。在定义问题时，首先需要结合现实情况，明确差距或矛盾所在，从而挖掘出有意义、有价值的问题。但矛盾是普遍存在的，世间万物也是普遍联系的，在解决一个具体问题时，无法考虑到所有可能的影响因素。因此，在定义问题时需要做出取舍，通过边界划定确定需要考虑的主要相关因素，而忽略那些我们认为（或假设）不重要的因素。其次要对问题进行明确的、可量化的描述，需要将非量化的描述词汇转化为具有确定标准的可量化指标，例如将“汽车销量高”转化为“汽车销售量超过 100 万台”。只有将问题定义清楚，才能有针对性地开展后续的数据分析流程，从而得出合理的结论。

4. 数据预处理是指综合运用数据清理、数据集成、数据归约、数据变换等多种处理方法，将各种原始数据加工成人们需要的标准的、干净的数据的过程。

四、实训题

略

项目二　数据描述性分析

一、选择题

1. B　　2. B　　3. A　　4. A　　5. C　　6. D

二、填空题

1. 数据的中心位置
2. 数据取值分散性
3. 直方图与茎叶图
4. χ^2 检验法
5. 多维观测数据

三、简答题

1. 数据描述性分析是对数据进行整理和统计，通过计算一些统计量如平均数、中位数、众数、方差等来描述数据的基本特征和分布情况的过程。

2. 数据描述性分析的主要目的是理解和解释数据的特征，包括数据的中心趋势、离散程度、偏度和峰度等，以便于进一步的数据分析、预测和决策。

3. 常用的数据描述性统计量包括平均数（均值）、中位数、众数、方差、标准差、偏度和峰度等。

四、实训题

略

项目三　线性回归分析

一、选择题

1. A　　2. A　　3. D　　4. B　　5. B　　6. C　　7. B　　8. C

二、填空题

1. 因变量　　自变量
2. 直线回归方程
3. 标准化残差直方图
4. 关系数检验、标准误差检验、F 检验、t 检验
5. 随机误差

三、简答题

1. 假设 1：

解释变量 x 是确定性变量，y 是随机变量。

假设 2：

随机误差项 ε 具有零均值、同方差和不序列相关性：

$$E(\varepsilon_i)=0 \qquad i=1,2,\cdots,n$$

Var （ε_i)= ^{2} i=1,2,⋯,n

Cov（ε_i, ε_j)=0 $i\neq j$ i,j= 1,2,⋯,n

假设 3：

随机误差项 ε 与解释变量 x 之间不相关：

Cov（X_i, ε_i)=0 i=1,2, ⋯,n

假设 4：

ε 服从零均值、同方差、零协方差的正态分布

ε_i~N（0, ^{2}) i=1,2, ⋯,n.

2. 解析：

i	1	2	3	4	5
x_i	2	3	4	5	6
y_i	2.2	3.8	5.5	6.5	7.0
x_iy_i	4.4	11.4	22.0	32.5	42.0
x_i^2	4	9	16	25	36
$\overline{x}=4$， $\overline{y}=5$， $\sum_{i=1}^{5}x_i^2=90$， $\sum_{i=1}^{5}x_iy_i=112.3$					

$$b=\frac{\sum_{i=1}^{5}x_iy_i-5\overline{x}\,\overline{y}}{\sum_{i=1}^{5}x_i^2-5\overline{x}^2}=\frac{112.3-5\times4\times5}{90-5\times4^2}=1.23.$$

于是，

$a=\overline{y}-b\overline{x}=5-1.23\times4=0.08$.

∴线性回归方程为： $\hat{y}=bx+a=1.23x+0.08$

3. 模型中一旦出现异方差，如果仍采用最小二乘法估计模型参数，会产生一系列不良后果：①参数估计量是无偏的、一致的，但不具有有效性。原因是，在证明无偏性和一致性时未用到同方差的假定，但是在证明有效性时用到了同方差假定。②参数估计量的方差出现偏误，变量的 t 检验和 F 检验失效。③异方差将导致预测区间偏大或偏小，预测失效。

四、实训题

略

项目四　对比分析

一、选择题

1. B　2. B　3. B　4. A　5. C　6. D　7. D

二、填空题

1. 绝对数　构成的百分比　环形图
2. 为蜘蛛图
3. 箱线图
4. 两个变量之间
5. 内部子总体
6. 对照参考的指标在所研究
7. 主观赋权法

三、简答题

1. （1）内部对比分析。内部对比分析是把数据的内部子总体交叉分类，对比不同类型数据的差别，从而发现问题。用于分类的定性变量又被称为“维度”，也就是说从不同维度分析对比，发现问题点，寻找影响因素。

（2）外部对比分析。外部对比分析是指对照参考的指标在所研究的数据之外，需要从其他渠道寻找补充。第一章曾经提到的对比分析有期间比较、实体比较、口径比较、结构比较，这些都是外部对比常用的方法，另外还可以和一些经验标准相比较，如流动比率和 2 比较，考试分数和 60（及格线）比较等。

（3）通过钻探寻找问题的原因。通过分类钻探，可以针对发现的问题寻找原因或关键点，从而有利于问题的解决。

2. 综合评价可以使用模糊数学的模糊综合评价方法，或者使用多元统计分析的因子分析方法，即计算因子得分，把因子得分加权平均，得到综合得分，进而依此排名。本书所述的常规综合评价方法是不涉及模糊数学、多元统计分析等其他学科知识的综合评价方法，也是目前在实际评价中应用较为广泛的一种方法。通常，常规综合评价方法需要解决好四个方面的问题：第一，综合评价指标体系的确定，对此应综合考虑评价的目的、价值取向，选取最合适的指标，以满足指标体系的系统性、科学性、可操作性等要求；第二，单指标无量纲化处理，即将不同量纲、不同性质的指标数值转换成单项得分，以利于指标之间的综合计算；第三，指标权重的确定，即根据每个评价指标在评价中的重要程度给予不同的权重，以进行不同指标单项得分的加权平均；第四，单项指标得分的加权平均，即运用一定的综合方法将不同指标得分进行汇总，以比较不同单位之间的得分排序。

3. 数据报告的内容一般包括描述分析、问题发现、对策建议、未来展望等。

数据分析报告曾有一个约定俗成的格式：一说明情况，二分析问题，三给出建议。

后来，又出现了另一种三段式：提出问题，分析问题，解决问题。

统计分析报告主要就三句话：是什么，为什么，怎么办。

严格说来，数据分析报告的结构应当根据分析的题材、内容、种类的不同而有所不同，即使是同一题材、同一内容、同一种类的分析报告，其格式也不能一成不变。但所有的数据分析

报告，都要力求做到有数字、有情况、有分析、有建议；而且数字要准确，情况要真实，分析要透彻，建议要可行。

四、实训题

略

项目五　聚类分析

一、选择题

1. B　　2. A　　3. A　　4. A　　5. C

二、填空题

1. 广义欧氏距离

2. 最短距离法　最长距离法　中间距离法　重心法　类平均法　可变类平均法　可变法　离差平方和法。

3. 夹角余弦　相关系数

4. 间隔尺度　顺序尺度　名义尺度

5. 样品　初始分类

6. 欧氏距离

三、简答题

1. 快速聚类分析是一个不断迭代的过程，其基本原理和迭代步骤如下：

（1) 首先需要用户指定聚类成多少类（比如 *K* 类)。

（2）然后 SPSS 确定 *K* 个类的初始类中心点。SPSS 会根据样本数据的实际情况，选择 *K* 个由代表性的样本数据作为初始类中心。初始类中心也可以由用户自行指定，需要指定 *K* 组样本数据作为初始类中心点。

（3）计算所有样本数据点到 *K* 个类中心点的欧氏距离。SPSS 按照距 *K* 个类中心点距离最短原则，把所有样本分派到各中心点所在的类中，形成一个新的 *K* 类，完成一次迭代过程。

（4） SPSS 重新确定 *K* 个类的中心点。SPSS 计算每个类中各个变量的变量值均值，并以均值点作为新的类中心点。

（5）重复上面的两步计算过程，直到达到指定的迭代次数或终止迭代的判断要求为止。

2. 系统聚类方法的主要类型包括最短距离法、最长距离法、类平均法、重心法、中间距离法、可变类平均法、可变法和离差平方和距离法（Ward 法）等。其中最短距离法和最长距离法都是以类与类之间的距离作为聚类的依据，类平均法则是计算两个类中所有样本之间的距离的平均值作为类与类之间的距离，重心法则是将类看作一个点，计算两个类重心之间的距离作为类与类之间的距离，中间距离法和离差平方和法则分别计算两个类中样本之间的中间距离和离差平方和作为类与类之间的距离。可变法在聚类过程中允许类别中变量的权重或重要性发生变化。这意味着不同的变量在聚类过程中可能会起到不同的作用，根据它们在聚类中的贡献程度进行调整。该方法可能更适合处理具有不同重要性的变量，或者在聚类过程中变量的重要性可能发生变化的情况。Ward 法是一种基于最小方差原理的聚类方法。它将两个类别的合并看作是

一次新类别的生成，新类别包含了原来两个类别的所有观测值。Ward 法的主要优点是能够处理数据集中的异常值，并降低它们对聚类结果的影响。这是因为该算法使用平均变量之间的平方和（SSE）作为距离度量方式，将新类别与原类别的距离定义为变量内部的差异程度。

项目六　主成分分析

一、选择题

1. B　　2. B　　3. C　　4. D　　5. A

二、填空题

1. 主分量分析
2. p 维空间中椭球体的主轴
3. 原始变量
4. 坐标系旋转
5. 原有变量
6. 0～1

三、简答题

1. （1）基本思路：数据标准化：首先，对原始数据进行标准化处理，即减去均值并除以标准差，以消除量纲和数量级对数据分析的影响。

计算协方差矩阵：计算标准化后数据的协方差矩阵，以获取各个特征之间的相关性信息。

特征值分解：对协方差矩阵进行特征值分解，得到特征值和对应的特征向量。这些特征向量表示了新的坐标系的方向，而特征值则表示了数据在该方向上的方差大小。

选择主成分：根据特征值的大小选择主成分。通常选择特征值较大的前几个主成分，因为它们包含了数据中大部分的信息。

转换坐标系：将原始数据从原始坐标系转换到由选定的主成分构成的新坐标系中，得到降维后的数据。

（2）主要应用：①特征提取与降维：主成分分析可以有效地提取数据的主要特征，去除冗余信息，从而简化问题的复杂性。这在图像处理、语音识别、机器学习等领域有着广泛的应用。②数据压缩：通过保留较多的主成分，可以在一定程度上减小数据的存储空间和计算负担，提高数据处理的效率。因此，主成分分析也常用于数据压缩。③可视化分析：对于高维数据，我们通常很难直接进行可视化。主成分分析可以将高维数据映射到二维或三维空间中，从而帮助我们直观地理解数据的结构和关系。④异常值检测：在主成分分析的结果中，远离主成分的数据点可能被视为异常值。因此，主成分分析也可以用于异常值检测。⑤数据预处理：在机器学习和数据挖掘中，主成分分析常被用作数据预处理的一种手段，以提高后续算法的性能和效率。

2. 由协方差矩阵和由相关系数矩阵出发进行主成分分析的区别主要体现在处理数据的方式和结果上。

首先，从协方差矩阵出发进行主成分分析时，其结果会受到变量单位的影响。主成分会倾向于多归纳方差大的变量的信息，对于方差小的变量就可能体现得不够，这就存在“大数吃小数”的问题。特别是当各指标之间的数量级相差悬殊，或各指标有不同的物理量纲时，直接由

协方差矩阵出发进行主成分分析可能会得到不准确的结果。

而由相关系数矩阵出发进行主成分分析时，数据通常会被先标准化。标准化后的数据消除了量纲和数量级的影响，使得主成分分析更加合理和准确。在这种情况下，主成分将更加公平地对待所有变量，不会偏向于方差大的变量。

其次，从协方差矩阵出发得到的主成分和从相关系数矩阵出发得到的主成分一般是不相同的。实际中，这种差异有时可能很大。由协方差矩阵出发求解主成分所得的结果与由相关系数矩阵出发求解主成分所得的结果在解释原始变量方差比例和主成分表达式上均有显著差别，且两者之间不存在简单的线性关系。

因此，选择从协方差矩阵还是相关系数矩阵出发进行主成分分析，需要根据实际的数据情况和研究需求来决定。如果各指标之间的数量级相差悬殊或各指标有不同的物理量纲，那么采用由相关系数矩阵出发进行主成分分析可能更为合适。如果数据已经经过适当的预处理，且各指标的量纲和数量级相对一致，那么直接从协方差矩阵出发进行主成分分析也是可以的。

3. 略

四、操作题

略

项目七　相关分析

一、选择题

1. C　　2. B　　3. C　　4. B　　5. D　　6．D　　7. A

二、填空题

1. 确定性　内在关联　相关程度　相关方向　相关形式
2. 一元相关　多元相关
3. 变量之间相关关系
4. 序变量、定序变量、顺序变量
5. 相应分析、关联分析、R-Q 型因子分析
6. 简单对应分析　多重对应分析　均值对应分析

三、简答题

1. 相关分析是研究两个或多个变量之间是否存在某种关联或依存关系的一种统计分析方法。它可以帮助我们了解变量之间的相关程度、方向和形式。

2. 相关系数是一种用于量化变量之间相关程度的统计分析指标。它的取值范围在−1 到 1 之间，其中 1 表示完全正相关，−1 表示完全负相关，0 表示无相关。相关系数可以帮助我们了解变量之间的相关程度和方向，从而判断它们之间的关系是否紧密。

3. 可以通过计算相关系数来判断两个变量之间是否存在相关关系。如果相关系数的绝对值接近 1，则说明两个变量之间存在较强的相关关系；如果相关系数的绝对值接近 0，则说明两个变量之间的相关关系较弱或无相关关系。

4. 正相关是指两个变量之间的变化方向相同，即当一个变量增加时，另一个变量也增加；

当一个变量减小时，另一个变量也减小。例如，身高和体重之间通常存在正相关关系。负相关是指两个变量之间的变化方向相反，即当一个变量增加时，另一个变量减小；当一个变量减小时，另一个变量增加。例如，温度和空调耗电量之间通常存在负相关关系。

项目八　时间序列分析

一、选择题

1. D　　2. A　　3. A　　4. A　　5. D

二、填空题

1. 时间点上　　数据
2. 时间推移　　平均值
3. 局部平稳
4. 长期趋势　　季节变动　　循环变动　　不规则变动

三、简答题

1. ARIMA 模型算法的基本思想可以概括为以下几个关键点：

（1）模型定义：

ARIMA 模型全称为差分自回归移动平均模型（Autoregressive Integrated Moving Average Model），记为 ARIMA（p,d,q）。

其中，AR 代表自回归（Autoregressive），p 是自回归项数；MA 代表移动平均（Moving Average），q 是移动平均项数；I 代表差分（Integrated），d 是时间序列成为平稳时所做的差分次数。

（2）基本思想：

将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列。

模型一旦被识别后，就可以从时间序列的过去值及现在值来预测未来值。

（3）模型组成：

自回归（AR）：描述当前值与历史值之间的关系，用变量自身的历史时间数据对自身进行预测。自回归模型的阶数 p 表示过去观测值的个数，即自回归系数的个数。

差分（I）：为了消除时间序列数据的非平稳性。差分过程是指对时间序列数据进行一阶或多阶的差分运算，使其转化为平稳的时间序列数据。差分的阶数 d 决定了进行几次差分操作。

移动平均（MA）：关注自回归模型中的误差项的累加。移动平均模型的阶数 q 表示过去观测值的误差个数，即移动平均系数的个数。

（4）模型建立过程：

模型识别：确定 ARIMA 模型的阶数（p, d, q）。可以通过查看自相关图（ACF）和偏自相关图（PACF）来初步判断。

参数估计：使用最大似然估计或其他方法对 ARIMA 模型的参数进行估计。

模型检验：对模型进行残差分析，判断模型是否符合时间序列数据的特征。常见的检验方法包括残差自相关系数检验、残差平稳性检验等。

（5）预测应用：

在模型通过检验后，可以利用该模型对未来的时间序列数据进行预测。预测的方法包括一步预测和多步预测等。

总结来说，ARIMA 模型算法的基本思想是通过识别时间序列数据中的自回归成分、移动平均成分和随机误差项，构建一个差分自回归移动平均模型，用于对时间序列数据进行建模和预测。

2. Prophet 模型算法的基本思想可以分为以下几个要点进行简述：

（1）模型概述：

Prophet 是 Facebook 在 2017 年开源的时间序列预测框架，适用于各种具有潜在特殊特征的预测问题，包括广泛的业务时间序列问题。

它对时间序列趋势变化点的检测、季节性、节假日以及突发事件具有更好的拟合效果。

（2）模型组成：

Prophet 模型是一个加法回归模型，由三个核心部分组成：trend（趋势项）、seasonality（季节项）及 holidays（假期项）。

趋势项用于捕捉时间序列的非周期性变化；季节项用于建模周期性变化，如每周、每年的季节性；假期项则用于建模节假日的影响。

（3）模型公式：

Prophet 模型的基本组成公式为： $y(t) = g(t) + s(t) + h(t) + \varepsilon_t$

$g(t)$是趋势函数，用于模拟时间序列值的非周期性变化。

$s(t)$表示周期性变化，包括季节性因素。

$h(t)$表示假期的影响，这些影响在一天或几天内以潜在的不规则时间表发生。

ε_t 是误差项，代表模型不适应的特殊变化，通常假设其服从正态分布。

（4）算法优势：

Prophet 算法支持时间序列的训练和预测，并且具有以下优势：

对缺失值或异常值的包容性强。

支持周期和趋势的多尺度性。

支持对法定假日或特殊日期的针对性训练。

（5）算法流程：

Prophet 模型内部由循环中的分析师与自动化两部分构成一个循环体系。

根据预测问题建立时间序列模型，对历史数据进行仿真，评估模型的效果。

根据出现的问题，进一步进行调整和建模，最终以可视化方式反馈整个预测结果。

（6）灵活性与可配置性：

Prophet 提供了大量可配置的参数，使用者可根据具体需求调整模型，比如调整季节性的拟合度、添加自定义节假日、添加自定义变量等。

（7）应用实践：

在实际应用中，Prophet 模型可以用于预测多种时间序列数据，如产品销售量、网站访问量、股票价格等。

综上所述，Prophet 模型算法的基本思想是通过建立一个包含趋势、季节性和节假日效应的加法回归模型，对历史时间序列数据进行拟合和预测，并提供了丰富的参数配置和灵活的模型

调整方式，以满足不同预测问题的需求。

3. （1）RNN（循环神经网络）的基本原理：

RNN 的核心在于其循环结构，这一结构允许信息在不同时间步之间传递。在每个时间步，RNN 接收当前的输入数据（如一个词的嵌入表示）和前一个时间步的隐藏状态，然后生成一个新的隐藏状态。这个新的隐藏状态不仅包含了当前时间步的信息，还融合了之前所有时间步的信息，因此 RNN 能够捕捉到序列数据中的上下文信息。RNN 中的权重参数是共享的，这确保了模型可以处理任意长度的序列数据。

（2）LSTM（长短期记忆网络）的基本原理：

LSTM 是为了解决 RNN 在处理长期依赖时的梯度消失和爆炸问题而设计的。它引入了三个门结构（输入门、遗忘门和输出门），以及一个细胞状态来控制信息的流动。遗忘门决定了哪些信息应该从细胞状态中丢弃，输入门决定了哪些信息应该被添加到细胞状态中，而输出门则控制细胞状态的输出。LSTM 的这种结构允许它学习并记住长期的依赖关系。

（3）GRU（门控循环单元）的基本原理：

GRU 是 LSTM 的一个简化版本，旨在保持 LSTM 效果的同时减少计算量。它只有两个门结构：重置门和更新门。重置门决定了如何将新的输入信息与先前的隐藏状态结合以产生候选隐藏状态，而更新门则决定了先前的隐藏状态有多少信息应该被保留以及有多少新的信息应该被包含进来。

（4）RNN、LSTM 与 GRU 的异同点：

① 相同点：

三者都是循环神经网络，旨在处理序列数据。

权重参数在时间步之间是共享的。

都具有记忆能力，能够捕捉序列数据中的长期依赖关系。

② 不同点：

结构：RNN 是基本的循环神经网络结构；LSTM 引入了门结构和细胞状态；GRU 是 LSTM 的简化版本，只有两个门结构。

能力：RNN 在处理长期依赖时存在梯度消失和爆炸的问题；LSTM 和 GRU 通过引入门结构有效地解决了这个问题，使得它们能够学习更长的序列。

参数和计算量：RNN 最简单，参数最少；LSTM 参数较多，计算复杂度高；GRU 的参数数量介于 RNN 和 LSTM 之间，计算速度较快。

（4）应用：在需要捕捉长期依赖关系且对计算量要求不高的场景下，LSTM 是较好的选择；而在对计算速度要求较高的场景下，GRU 可能是一个更好的选择。

四、操作题

略