

Lesson 3 Pre-Test

第三课摸底测试

Student/姓名: Scores/得分:

English	Answer
What is a CSV file?	
a CSV is a method of JavaScript Object Notation.	
b CSV files are rows of data or values separated by commas.	
c CSV makes data readily available for analytics, dashboards, and reports.	
d CSV files are a standard way to store data across platforms.	
What are residuals?	
a Residuals are a method for handling identified outliers.	
b Residuals are data removed from the dataframe.	
c Residuals are the difference between the actual values and the values predicted by a given model.	
d Residuals are a method to standardize data.	
If removal of rows or columns of data is not an option, why must we ensure that information is assigned for missing data?	
a Missing data may bias the dataset.	
b Most models will not accept blank values in our data.	

中文	答案
什么是 .csv 文件?	
a CSV 是 JavaScript Object Notation 的一种方法。	
b CSV 文件是由逗号分隔的数据行或值。	
c CSV 使数据易于用于分析、仪表板和报告。	
d CSV 文件是跨平台存储数据的标准方式。	
什么是残差?	
a 残差是一种处理已识别异常值的方法。	
b 残差是从数据框中删除的数据。	
c 残差是实际值与给定模型预测值之间的差异。	
d 残差是一种标准化数据的方法。	
如果删除数据行或列不是一种选择, 为什么我们必须确保为丢失的数据分配信息?	
a 缺失数据可能会使数据集产生偏差。	
b 大多数模型不会接受我们数据中的空白值。	

c	Assigning information for missing data improves the accuracy of the dataset.	
d	Information must be assigned to prevent outliers.	
What are the two main data problems companies face		
a	Data sampling and categorization	
b	Lack of training and expertise	
c	Lack of relevant data and bad data	
d	Outliers and duplicated data	
What is meant by the Messy Data?		
a	Duplicated or unnecessary data.	
b	Inconsistent text and typos.	
c	Missing data.	
d	All of the above.	
What is an outlier?		
a	Outlier is a data point that has the highest or lowest value in the dataset.	
b	Outlier is a data point that is very close to the mean value of all observations.	
c	Outlier is an observation in dataset that is distant from most other observations.	
d	Outlier is a data point that does not belong in our dataset.	
How do we identify outliers in our dataset?		
a	We can identify outliers only by calculating the minimum and maximum values in the dataset.	
b	We can identify outliers both visually and with statistical calculations.	
c	We can only identify outliers by using some statistical calculations.	
d	We can only identify outliers visually through building plots.	

c	为缺失数据分配信息可以提高数据集的准确性。	
d	必须分配信息以防止异常值。	
公司在开始使用人工智能/机器学习时面临的两个主要数据问题是什么？		
a	数据采样和分类	
b	缺乏培训和专业知	
c	缺乏相关数据和不良数据	
d	异常值和重复数据	
混乱的数据是什么意思？		
a	重复或不必要的数	
b	文字不一致和错别字。	
c	缺失数据。	
d	上述所有的。	
什么是异常值？		
a	异常值是数据集中具有最高或最低值的数据点。	
b	离群值是非常接近所有观测值的平均值的数据点。	
c	离群值是数据集中与大多数其他观测值相距甚远的观测值。	
d	异常值是不属于我们数据集中的数据点。	
我们如何识别数据集中的异常值？		
a	我们只能通过计算数据集中的最小值和最大值来识别异常值。	
b	我们可以通过视觉和统计计算来识别异常值。	
c	我们只能通过一些统计计算来识别异常值。	
d	我们只能通过构建图直观地识别异常值。	