考点总结 - Statistics for Data Science Lecture 4

1. 探索性数据分析（EDA）步骤：浏览数据集，单独分析每个变量，研究变量之间的关系，最后通过图表和数值总结呈现数据。

2. 数据集案例：包括空气质量数据集（纽约，1973年）和钻石价格数据集（54,940行数据，10个描述性变量）。

3. R工具和包：DataExplorer、dlookr等，用于加速数据探索和准备分析。

# 考点总结 - Statistics for Data Science Lecture 5

1. 推断统计学：从样本数据推导出关于总体参数的结论。

2. 估计类型：包括点估计和区间估计，点估计是用于近似总体参数的单个数值。

3. 估计量的性质：包括无偏估计量（其偏差为零）和有偏估计量。

4. 如何选择一个好的估计量：考虑无偏性、一致性和有效性等统计性质。

# 考点总结 - Statistics for Data Science Lecture 6

1. 推断统计学（假设检验）：通过样本数据推断总体参数，并进行假设检验来判断总体参数的有效性。

2. 假设检验的定义：假设检验用于检验总体参数的声明是否成立，通常通过计算p值来进行推断。

3. 实际应用问题：例如，通过抽样检验新生产工艺是否比旧工艺更可靠。

# 考点总结 - Statistics for Data Science Lecture 7

1. 统计功效（Power）：研究能够正确拒绝零假设的概率，避免Type II错误。

2. 效应量（Effect Size）：衡量处理效果的大小，常用于量化研究中的处理影响程度。

3. 影响统计功效的因素：样本大小、显著性水平（alpha）、效应量等。

# 考点总结 - Statistics for Data Science Lecture 8

1. 回归分析：用于预测或描述变量之间的关系，特别是用于定量数据之间的关系。

2. 统计学习与机器学习的区别：统计学习侧重于模型的解释性和精度，而机器学习侧重于大规模数据的应用和预测精度。

3. 回归模型的应用：例如，通过回归模型预测工资与年龄、教育年限等变量的关系。

# 考点总结 - Statistics for Data Science Lecture 9

1. 回归分析：用于描述和预测定量变量之间的关系，回归模型包括简单线性回归和多元线性回归。

2. 简单线性回归：考虑一个预测变量，用于描述单一因果关系。

3. 多元线性回归：考虑多个预测变量，适用于多个因素共同影响的情况。

4. 回归模型应用：例如，预测广告费用与销售额之间的关系。

# 考点总结 - Statistics for Data Science Lecture 10

1. K折交叉验证（K-fold Cross-validation）：提供了偏差和方差的平衡，能够更准确地评估模型性能。

2. 重复K折交叉验证：通过多次重复K折交叉验证来减少结果的方差。

3. 回归模型性能度量：包括均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）、R平方（R²）等。

4. 均方误差（MSE）与均方根误差（RMSE）的区别：MSE对异常值敏感，而RMSE易于解释，适用于非正态分布数据。

5. R平方（R²）：用于衡量回归模型的拟合优度，越高表示模型解释数据变异的能力越强。

# 考点总结 - Statistics for Data Science Lecture 10

1. 描述性统计：用于总结和描述数据的主要特征，如均值、方差、分布等。

2. 数据集的关键特征：在统计研究中，描述数据集时需要回答数据的描述、数据集大小、变量定义等核心问题。

3. 变量的定义：在统计研究中，变量是指数据集中可以变化的量，变量的定义在数据分析中至关重要。

# 考点总结 - Statistics for Data Science Lecture 2

1. 数据的定义：数据是收集的事实和统计信息，用于参考或分析。

2. 数据来源：区分原始数据（Primary Data）和二手数据（Secondary Data）。

3. 定类数据（Nominal Data）：这类数据是定性的，无法进行数学计算，主要用于分组和分类。