项目需求

1. 数据收集需求

定义所需数据类型，如客户购买历史、产品偏好、市场趋势等。

确定数据来源，包括内部数据（如销售记录、库存数据）和外部数据（如市场调研报告）。

2. 数据处理需求

需要对数据进行清洗、转换和整合，以确保数据质量。

确定数据处理流程和技术，如ETL（Extract, Transform, Load）。

3. 分析模型需求

确定分析目标，如客户细分、销售预测、市场趋势分析等。

选择合适的统计方法和算法，如回归分析、聚类分析、时间序列分析等。

4. 报告和可视化需求

设计报告格式和内容，确保信息呈现清晰、直观。

选择合适的数据可视化工具和技术，如图表、仪表板等。

5. 用户定制化需求

根据不同客户的需求，提供定制化的数据分析报告。

确定用户定制化的范围和限制。

6. 系统集成需求

确保数据分析模块能够与ERP系统中的其他模块（如CRM、库存管理）无缝集成。

7. 安全性和合规性需求

遵守数据保护法规，确保客户数据的安全和隐私。

8. 技术支持和培训需求

提供必要的技术支持和用户培训，帮助客户理解和使用数据分析结果。

项目计划

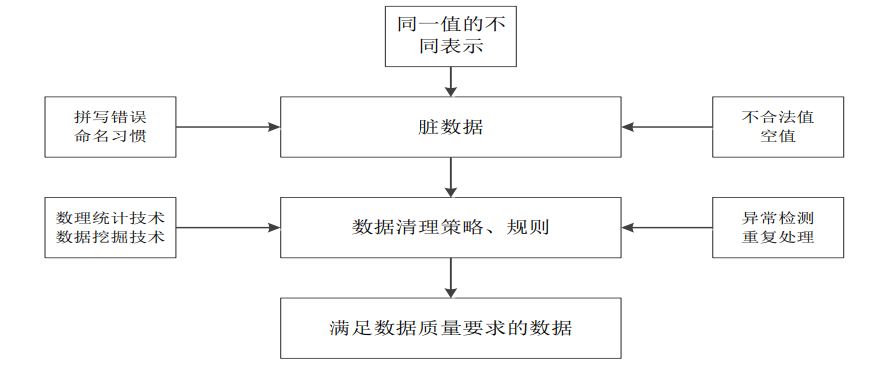
阶段一：数据收集与分析

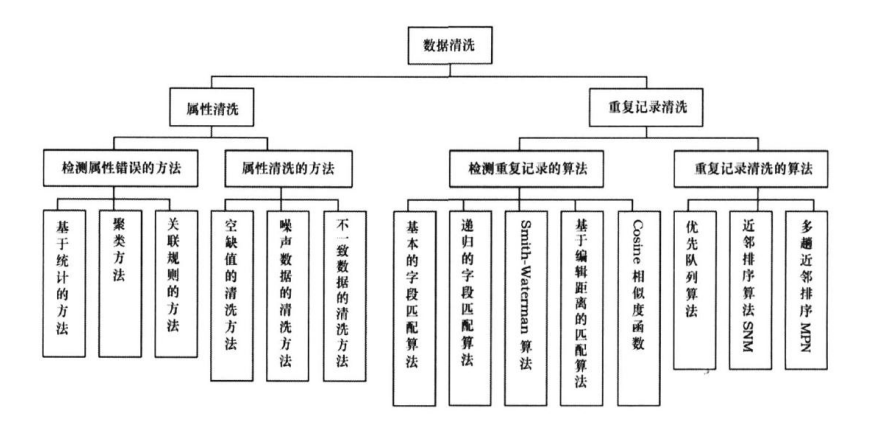
与客户沟通，明确数据分析的目标和需求。

确定数据收集的范围和方法。

实施数据收集计划，从ERP系统和其他来源获取数据。

进行数据清洗和预处理，确保数据质量。其中的数据间关联错误、数据本身错误、重复数据的冗余这些问题是影响单数据源的主要问题。误差是非常可能发生这种问题的原因。数据的关联错误一般是由人为造成的其中大多数为(1)数据的重复。(2)数据的本身错误。 (3)数据之间关联错误。使用数理统计、数据挖掘或预定义相对应领域清洗规则将脏数据或者说不符合要求的数据转化为满足要求的干净数据。



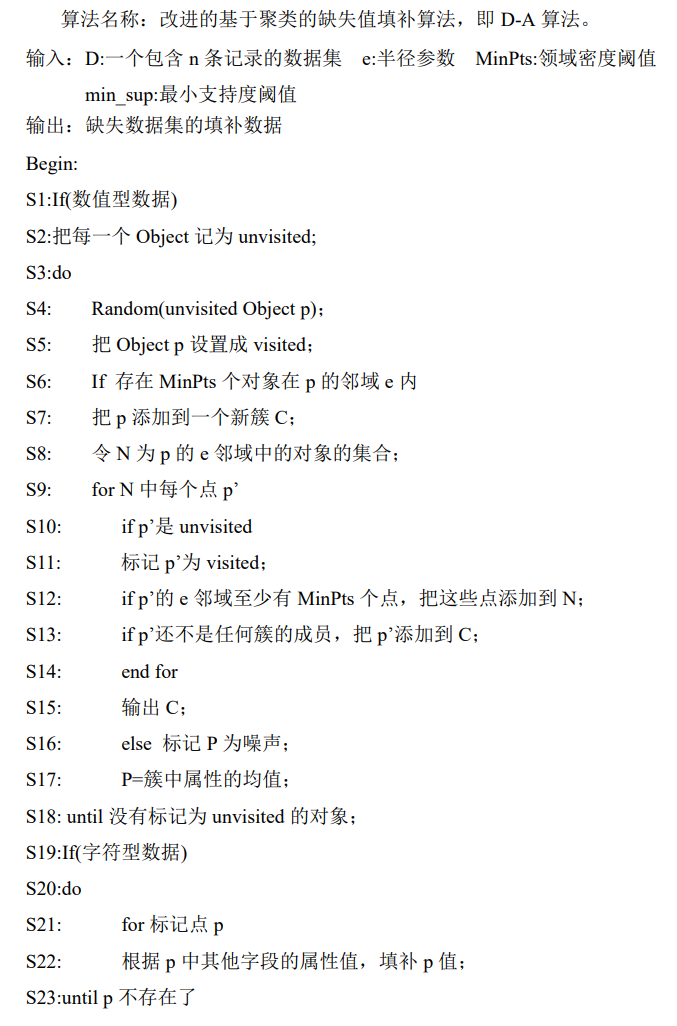


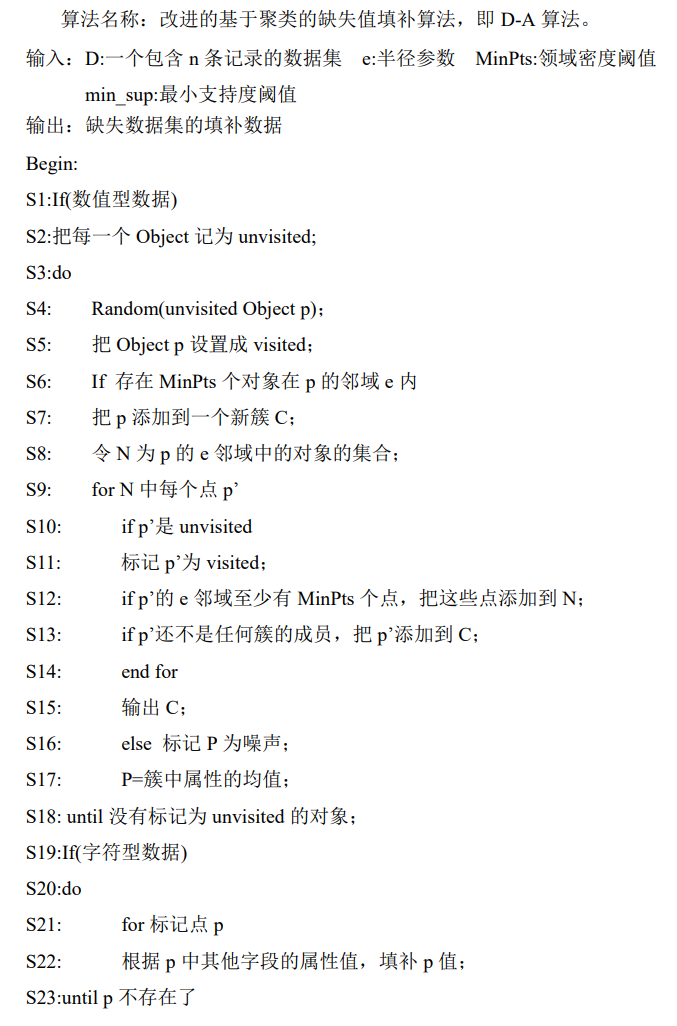
依据此我们设计了如下的数据清洗流程：

1. 数据分析阶段：为了发现那种错误和不一致要被删除，一种详细的数据分析是 需要的。除了数据和数据样本的人工检查，数据的字段内容的提取和发现查出数据问 题这些工作都是由指定程序来做的。
2. 定义工作流和转换规则：根据数据源的数量，和它们的异质程度以及数据的脏 乱程度，大量的数据转换和清洗步骤可以被执行。有时，一个模式的转换通常被用于 匹配一个公共数据来源模型；对于数据仓库来说，代表这一个典型的关系表示法。早 期的数据清洗不仅能满足单数据源的实例层的需要，下一步可以进行数据集成。后来 的步奏可以清洗模式层/数据集成和多数据源的实例层问题。对于数据仓库而言，对于 数据流转换的控制和清洗应该被特殊的制定到 ETL（用来描述数据从抽取，加载，转换到目的端的过程，常用于数据仓库中。也可由于其他方面）的流程当中。
3. 验证：一个转换模式的正确性和效率性，以及这种转换规则的定义应该被测试 和评估。应该有必要从源数据当中提取样本来提高定义的质量。多次迭代的分析，设 计和验证的步骤可能是必要的。例如，由于一些错误经过一些转换之后才变得明显。
4. 转型：通过加载和运行 ETL 工作流执行的转换步骤来刷新数据仓库或多个数 据源的查询回答。
5. 清洗后的回流：在（单数据源的）错误被删除后，干净的数据也替代原始的脏 数据回到数据源当中，为了避免遗留的应用程序重新为清洁工作进行数据的提取。

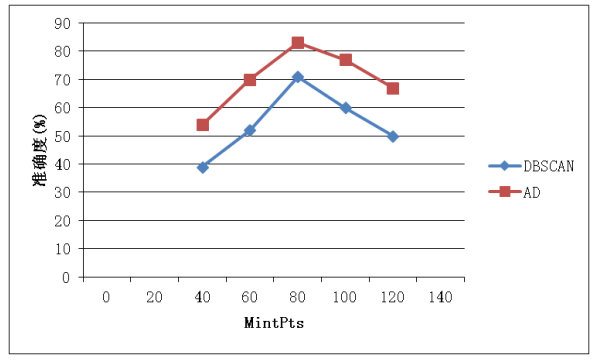
最后再根据清洗后获得的干净数据集的纯度来评价数据集。

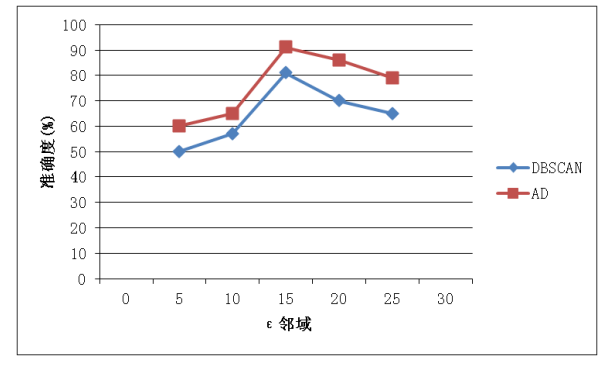
以下是数据清洗的伪代码：

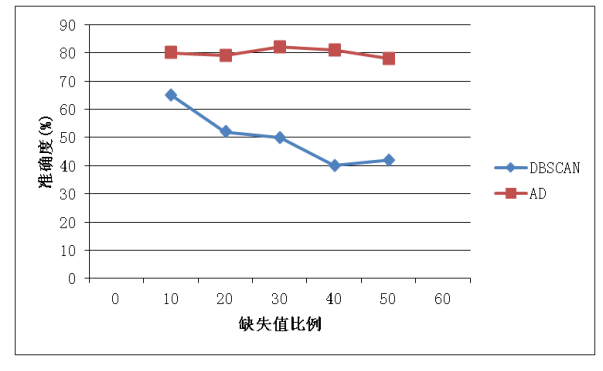




在待测数据集上采用 DBSCAN 和 D-A 算法分别在不同条件下对比了两种算法填 补准确度的情况。图 3.3 展示了在相同领域范围内 D-A 算法和 DBSCAN 算法明显在 填补准确度方面的情况。

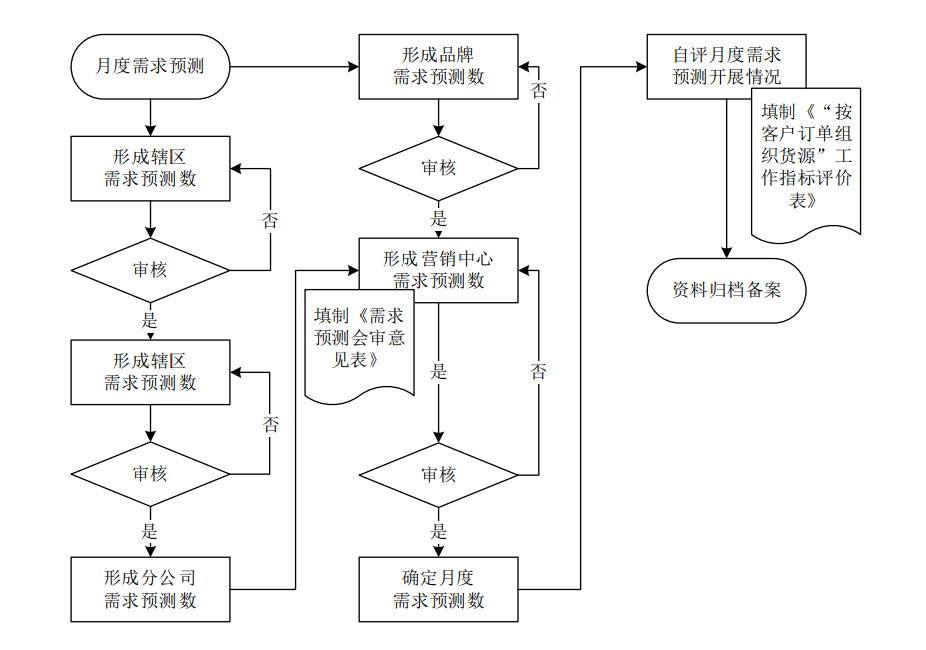


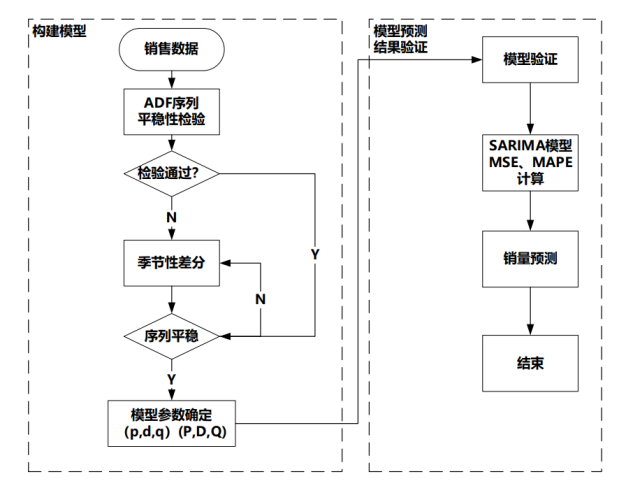




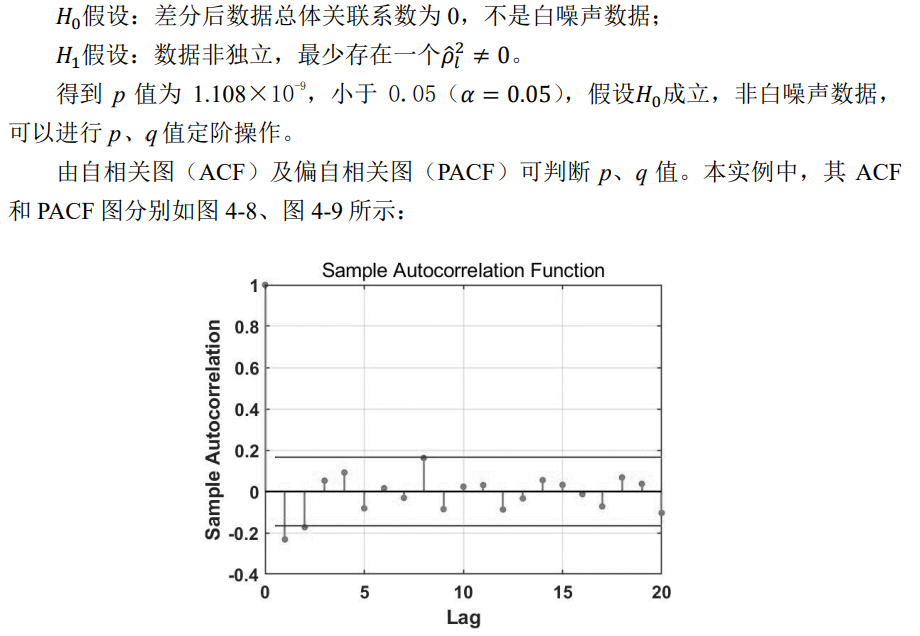
阶段二：分析模型开发

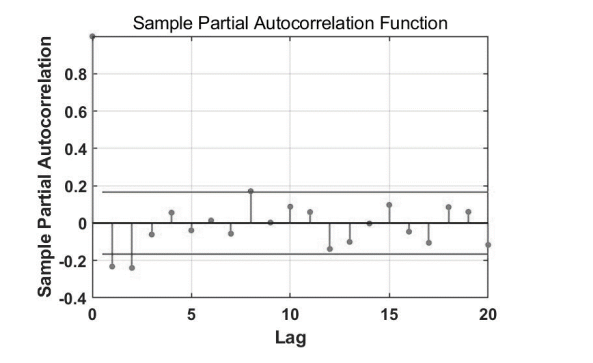
选择和开发适合的数据分析模型和算法，进行销量预测。

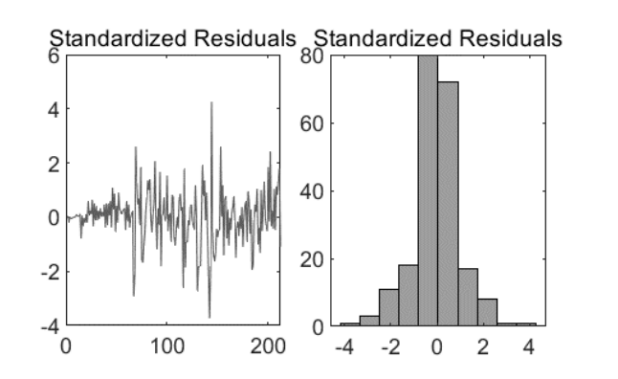


为此我方设计了一套基于SARIMA 模型的预测算法。

ADF 根检验假设：

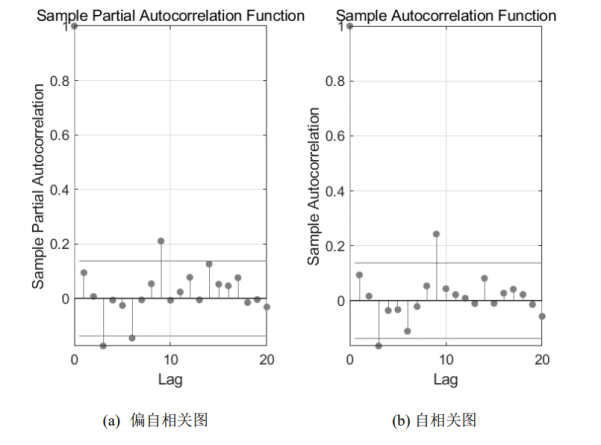




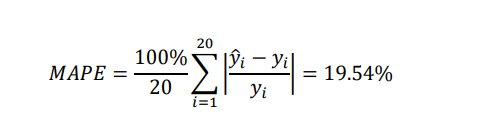


通过假设检验后进行初步分析，验证模型的有效性。

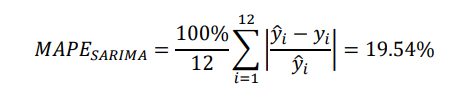
其中进行残差检验是最好的方式模型建模完毕之后，若建立的预测模型已经很好的提取了序列的有效信 息，则模型的残差序列没有相关性，那么残差序列就是白噪声，并在零均值以上呈正态分布。



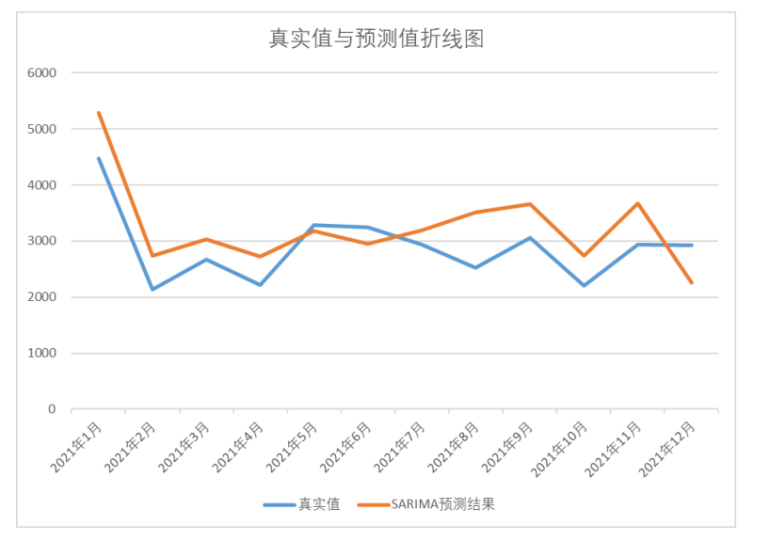
利用SARIMA模型预测公司销量数据与真实值对比。本司择取均方误差（MSE）和平均绝对百分比误差（MAPE）进行评估。



其平均绝对百分比误差为：



其对比结果如图：



阶段三：报告和可视化设计

设计数据分析报告的格式和内容。

开发数据可视化工具，如仪表板。

公司设计的终端显示系统是一个高度集成和实时更新的监控平台，它能够为管理层提供关键的业务洞察。系统包括以下几个核心图表：

1、海运陆运吨位占比图：实时展示不同运输方式的吨位比例，帮助管理层评估物流效率和成本。

2、各仓库发货量占比图：通过可视化的方式呈现各仓库的发货量，便于监控库存和发货效率。

3、各用户交易情况：展示用户交易的频率和金额，帮助识别关键客户和市场趋势。

4、各销售员销售额情况：实时追踪销售员的业绩，通过图表形式展示，以激励团队并指导资源分配。

此外，系统还具备以下特点：

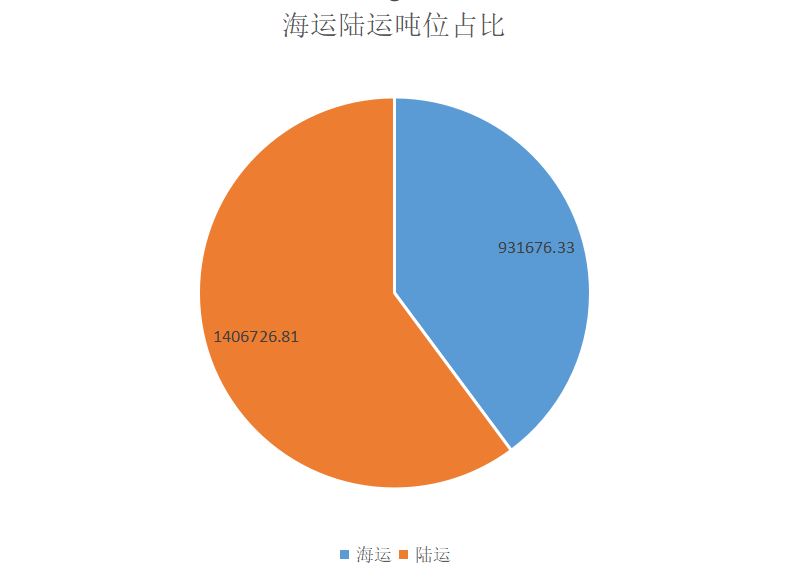
1、实时数据更新，确保信息的准确性和及时性。

2、交互式设计，允许用户深入探索特定数据点。

3、排序功能，使用户能够根据特定标准查看数据。

4、预警系统，自动提醒管理层关注异常指标。

一、海运陆运吨位占比图



通过对比海运陆运吨位数据，分析可得

1、销售分布： 哪些区域或客户更倾向于海运或陆运，了解产品销售的地理分布。

2、运输效率： 海运和陆运在运输时间、成本和可靠性方面的效率。

3、运输成本： 通过吨位数据，可以计算不同运输方式的单位运输成本，为成本控制提供依据。

4、市场需求： 吨位数据可以反映不同地区或不同时间点的市场需求强度。

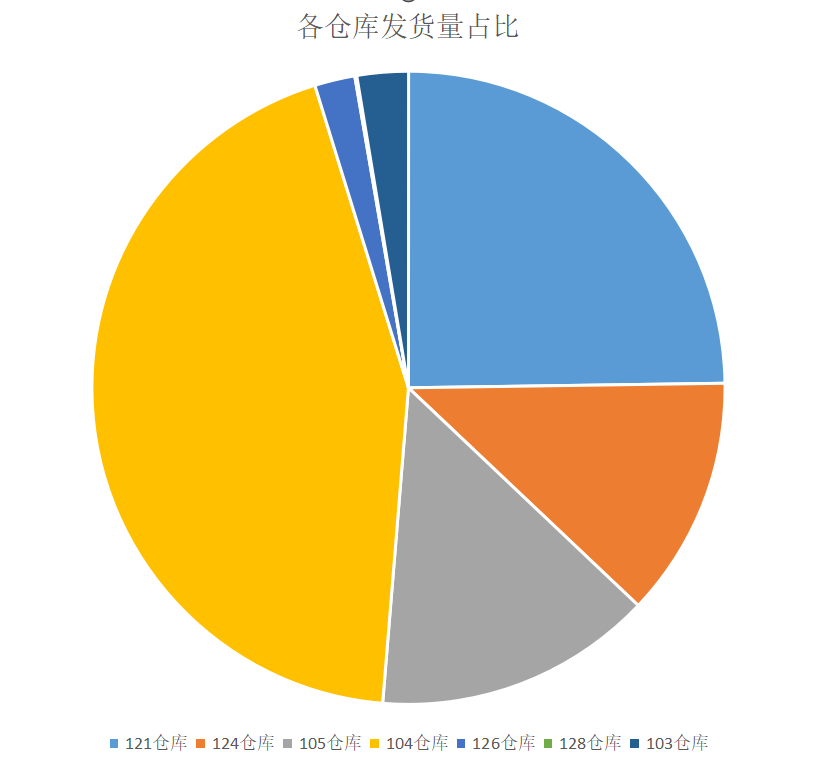
5、客户偏好： 分析客户对运输方式的选择，了解客户偏好和需求。

6、风险评估： 通过分析运输数据，评估供应链中的潜在风险点，如运输途中的损失或延误。

7、供应链优化： 识别供应链中的瓶颈和改进机会，优化整体供应链效率。

8、定价策略： 根据运输成本和市场情况，调整产品的定价策略。

二、各仓库发货量占比



1、地区销售动态： 发货量可以反映不同地区的销售活跃度和市场需求。

2、仓库使用效率： 了解哪些仓库的使用率高，哪些可能被低估，从而调整资源分配。

3、库存分布： 分析库存在各个仓库的分布情况，优化库存平衡。

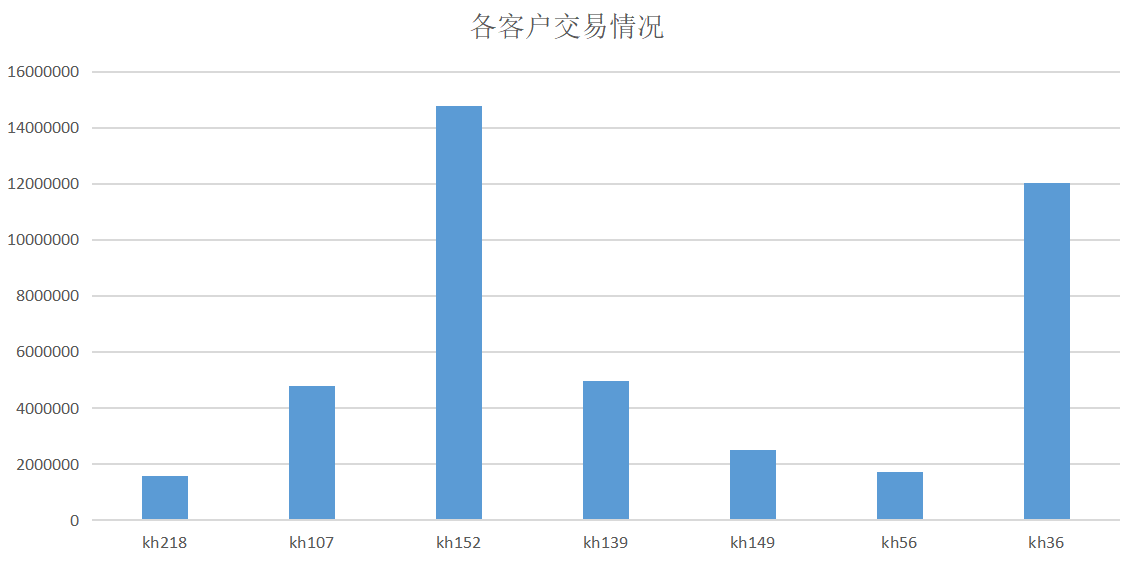
4、物流成本分析： 根据发货量，评估物流成本，包括运输和存储成本。

5、供应链优化： 识别供应链中的效率瓶颈，改进物流路径。

6、风险管理： 评估过度集中的仓库可能带来的风险，如自然灾害或运输中断。

7、投资回报分析： 分析各仓库的发货量与投资之间的关系，评估投资回报率。

三、各用户交易情况



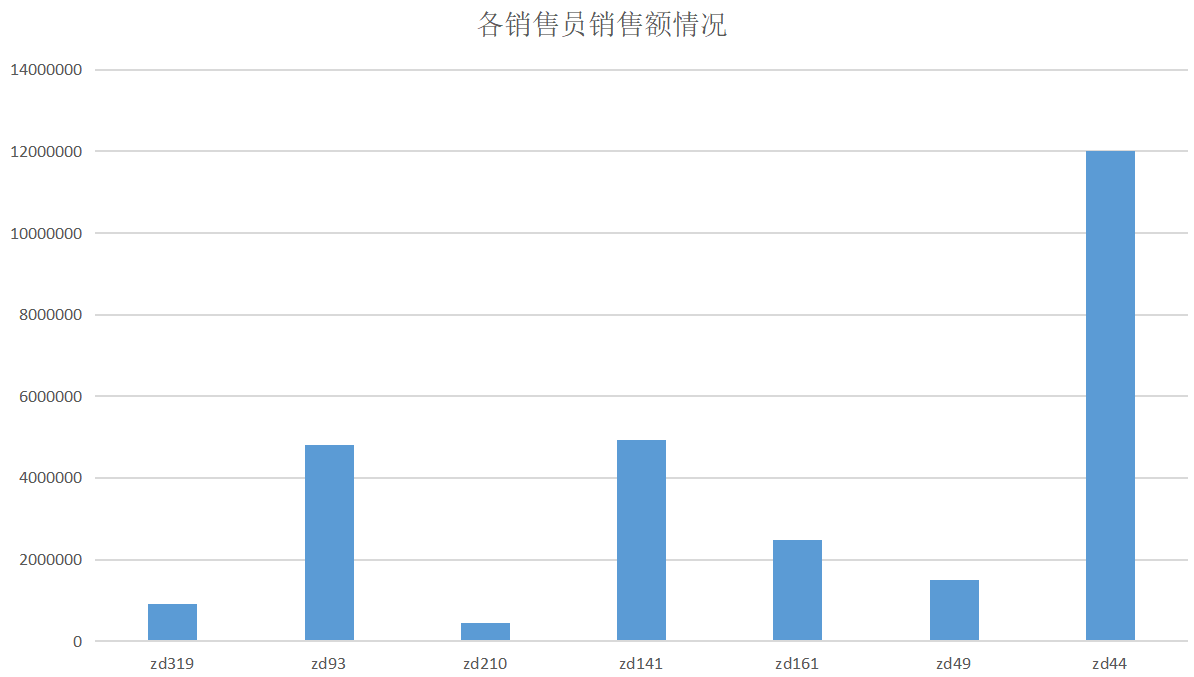
1、用户消费能力： 分析不同用户的交易额大小，了解他们的消费能力。

2、用户分层： 根据交易额将用户分为不同的层级，如高价值用户、中等价值用户和低价值用户。

3、收入和利润分析： 评估不同用户群体对企业收入和利润的贡献。

4、用户满意度： 高交易额可能表明用户对产品或服务的高度满意。

四、各销售员销售额情况



1. 个人销售能力： 评估每个销售员的销售技能和能力。
2. 业绩排名： 根据销售额对销售员进行排名，识别顶尖销售员。

3、销售员贡献度： 了解每位销售员对公司总销售额的贡献。

4、客户满意度： 销售员的销售额可能与其客户满意度相关联。

阶段五：系统集成和测试

将数据分析模块集成到ERP系统中。

进行全面测试，确保模块的稳定性和准确性。

阶段六：用户培训和反馈

为用户提供数据分析模块的培训。

收集用户反馈，进行必要的调整和优化。

阶段七：部署和上线

正式部署数据分析模块到生产环境。

监控系统性能，确保稳定运行。

阶段八：持续支持和维护

提供持续的技术支持和维护服务。

定期更新分析模型和报告，以适应市场变化。

阶段九：项目评估和优化

定期评估项目的效果和用户满意度。

根据评估结果进行项目优化和功能升级。

预期讨论目标

确定数据分析对提高产品销售的具体贡献。

讨论如何整合市场和客户数据以提供深入的洞察。

确定如何通过数据分析提高客户满意度和忠诚度。

讨论数据分析结果如何支持决策制定过程。