

**Oracle教育实训方案**

基于纽约出租车的大数据分析

**2018年4月**



目录

[综述： 3](#_Toc511744270)

[一、实训项目 3](#_Toc511744271)

[1、名称 3](#_Toc511744272)

[2、业务背景 3](#_Toc511744273)

[4、 数据字典 5](#_Toc511744274)

[5、实训方案概况 6](#_Toc511744275)

[6、实训需要讲授技术点 6](#_Toc511744276)

[二、实训开发环境 7](#_Toc511744277)

[1、软件环境 7](#_Toc511744278)

[2、硬件环境 7](#_Toc511744279)

[3、 开发语言 7](#_Toc511744280)

[三、 实训目标 7](#_Toc511744281)

[四、实训时间 8](#_Toc511744282)

[五、实训地点 8](#_Toc511744283)

[六、实训项目实施计划 8](#_Toc511744284)

[七、项目分析要点 9](#_Toc511744285)

[1、1-6月用车情况 9](#_Toc511744286)

[2、乘客数量分布 10](#_Toc511744287)

[3、支付方式 12](#_Toc511744288)

[4、平均距离及里程分布 13](#_Toc511744289)

[5、乘客叫车方式 15](#_Toc511744290)

[6、24小时使用趋势 17](#_Toc511744291)

[7、气温与出租车使用情况的关系 18](#_Toc511744292)

[八、项目分析步骤 20](#_Toc511744293)

[1、源数据和业务需求分析 20](#_Toc511744294)

[2、逻辑模型建立 21](#_Toc511744295)

[3、大数据平台建立 21](#_Toc511744296)

[4、物理模型创建 22](#_Toc511744297)

[5、ETL实现 23](#_Toc511744298)

[6、通过Hue图形化进行数据的分析和展现 23](#_Toc511744299)

[九、项目提交物 25](#_Toc511744300)

[十二、实训标准 25](#_Toc511744301)

[十三、实训学员日常管理制度 25](#_Toc511744302)

[十四、实训周报模板 27](#_Toc511744303)

[十五、实训考核制度及体系 28](#_Toc511744304)

[1、成绩得分 28](#_Toc511744305)

[2、项目结束评价 28](#_Toc511744306)

[3、团队项目成绩 29](#_Toc511744307)

[4、个人成绩 29](#_Toc511744308)

[5、实训最终成绩 30](#_Toc511744309)

# 综述：

甲骨文（oracle）公司作为世界500强的企业是目前唯一能为企业提供整体IT解决方案的公司，在50多种产品和技术领域处于领先地位，是技术标准的制定者。其拥有大量的企业用户，也最了解企业真正需求。为雇主企业提供优秀的技术支撑人才也就成为了甲骨文的迫切需要解决的问题。然而通过与高等院校的合作，成为其为企业提供人才储备的重要途径之一。

# 一、实训项目

## 1、名称

纽约出租车大数据分析

## 2、业务背景

在纽约，出租车分为两类：黄色和绿色。黄色出租(Yellow TAXI)车可以在纽约五大区（布朗克斯区、布鲁克林区、曼哈顿、皇后区、斯塔滕岛）内任何地点搭载乘客。绿色出租车(Green TAXI)则被规定只允许在上曼哈顿、布朗克斯区、皇后区和斯塔滕岛接客，这两类出租车均由私人公司经营并受到纽约市出租车和轿车委员会（NYC Taxi and Limousine Commission）的监管。本章对出租车1月——6月的数据进行分析，探究出租车的是使用趋势，用户使用习惯以及天气因素对出租车使用量的影响。

黄色和绿色出租车活动区域图如下：



绿色出租车可以在绿色区域自由载客，在灰色区域（肯尼迪国际机场和拉瓜迪亚机场）只能预约载客，在黄色区域是不能上客的。正因为这种营业区域的限制，造成了下面我们看到的很多有意思现象。  
**3、数据源地址**  
 <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml>

## 数据字典

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **名称** | **数据类型** | **数据格式** |
| 1 | [出租等级](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/915) | string | 文本字符串格式 |
| 2 | [执照号](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/916) | string | 文本字符串格式 |
| 3 | [提供商ID](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/917) | string | 文本字符串格式 |
| 4 | [比率代码](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/918) | string | 文本字符串格式 |
| 5 | [出租车挂靠标识](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/919) | string | 文本字符串格式 |
| 6 | [上车时刻](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/920) | string | 时间格式 YYYY/MM/dd |
| 7 | [下车时刻](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/921) | string | 时间格式 YYYY/MM/dd |
| 8 | [乘客数量](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/922) | int | 普通整数格式 |
| 9 | [行驶时长](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/923) | int | 普通整数格式 |
| 10 | [行驶距离](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/924) | float | 普通浮点数格式 |
| 11 | [上车位置经度](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/925) | float | 普通浮点数格式 |
| 12 | [上车位置维度](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/926) | float | 普通浮点数格式 |
| 13 | [下车位置经度](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/927) | float | 普通浮点数格式 |
| 14 | [下车位置维度](http://dataju.cn/Dataju/web/metaDataForAttributeDetail/928) | float | 普通浮点数格式 |

## 5、实训方案概况

* 对纽约出租车源数据和业务需求进行分析；
* 使用Hdfs作为数据存储区域，在Hdfs的基础上通过建模构建Hive数据仓库；
* 通过Flume来采集纽约出租车数据到数据仓库的ODS层；
* 通过Mapreduce来转码ODS层数据到数据仓库的基础层；
* 通过HiveSQL汇总数据到数据仓库的汇总层；
* 通过HiveSQL生产应用层数据；
* 通过Hue图表展现应用层数据；

## 6、实训需要讲授技术点

* 大数据技术（Hdfs、Mapreduce、Flume、Hive、Hue和yarn等技术）
* 数据仓库原理
* 仓库模型建立
* 元数据管理
* ETL技术
* 数据分析和展现

# 二、实训开发环境

## 1、软件环境

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 版本 | 备注 |
| 1 | Ubuntu系统 | 16.04 | 免费版 |
| 2 | JAVA虚拟机 | 1.7以上 | Oracle官网下载 |

## 2、硬件环境

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 名称 | 数量 | 备注 |
| 1 | Pc服务器 | 4台以上 | 每台配置要求：内存32G+、CPU主频2.5G+、硬盘300G+ |
| 2 | 局域网 | 通畅 |  |
| 3 |  |  |  |

## 开发语言

JAVA 、Python、Scala均可

# 实训目标

1、 通过实训能够使同学们了解大数据技术和BI的实施过程(主要包括需求和源数据分析、dw和dm建模、ETL过程、数据分析和图表展现)。

2、训练学生的思维能力、自学能力和trouble-shooting能力。

# 四、实训时间

2周

# 五、实训地点

机房

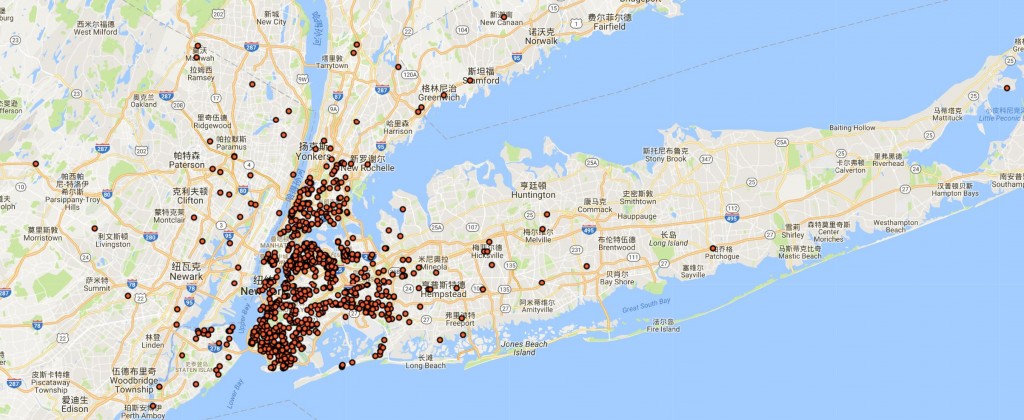
# 六、实训项目实施计划

|  |  |  |  |
| --- | --- | --- | --- |
| **序号** | **时间** | **内容** | **实现目标** |
| **第一阶段：基础知识储备** | | | |
| 1 | 第一周  第1天 | Linux命令学习  JAVA基础知识学习 | Linux常用命令讲解和学习  学习（复习）JAVA基础知识和语法 |
| 2 | 第一周  第2天 | Hdfs学习  通过API访问Hdfs | Hdfs原理和命令学习  通过Java API来读写Hdfs上的数据 |
| 3 | 第一周  第3天 | Mapreduce学习  Flume学习 | Mapreduce原理介绍，并编写第一个Mapreduce程序  Flume介绍，并能通过flume进行数据采集和处理 |
| 4 | 第一周  第4天 | Hive学习  数据仓库介绍 | 能通过Hive进行数据分析  学习数据仓库相关知识 |
| 5 | 第一周  第5天 | 仓库模型建立  图形化分析工具介绍 | 介绍仓库模型建立方法论  介绍HUE图形化分析工具 |
| **第二阶段：实施阶段，数据分析阶段** | | | |
| 6 | 第二周  第1天 | 大数据平台搭建 | 搭建大数据平台 |
| 7 | 第二周  第2天 | 数据仓库模型建立 | 数据仓库模型建立 |
| 8 | 第二周  第3天 | 数据采集入库 | 采集纽约出租车数据到数据仓库 |
| 9 | 第二周  第4天 | 分析 | 详细数据分析，并用图表对分析结果进行展示 |
| 10 | 第二周  第5天 | 项目收尾  项目说明书  项目评审 | 使用手册编写  部署文档书写  项目总结文档  PPT准备  项目评审 |

# 七、项目分析要点

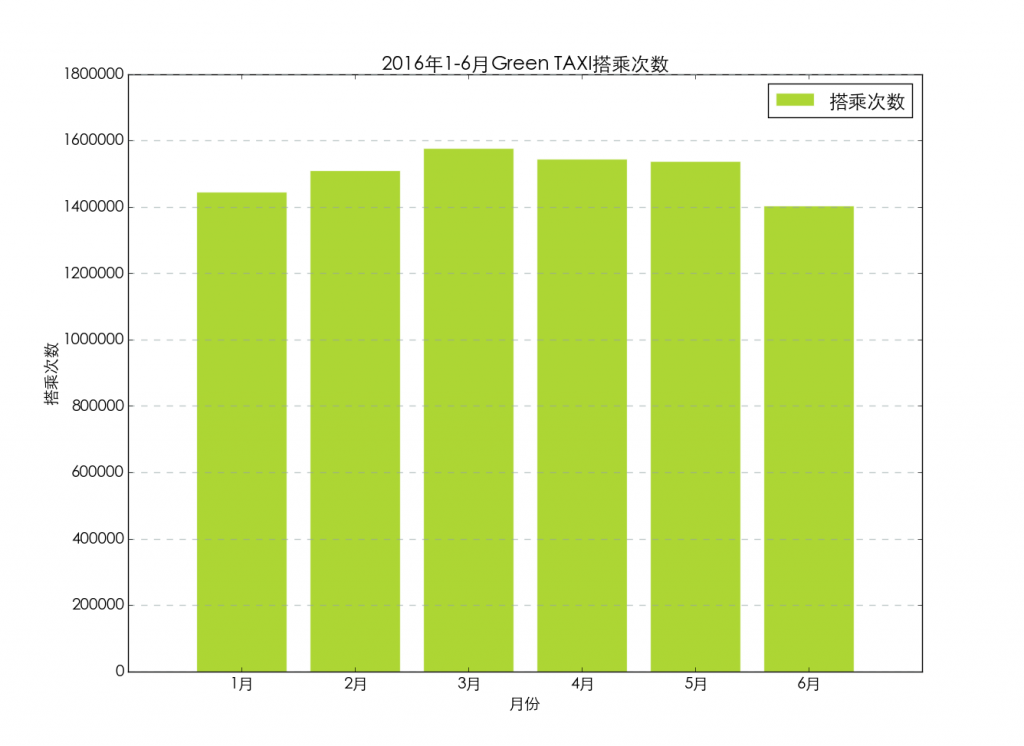
## 1、1-6月用车情况

这是一张实时Green TAXI的位置分布图。Green TAXI可以从曼哈顿北部（西110街和东96街的北部），布朗克斯，皇后区（不包括机场），布鲁克林和史坦顿岛的街道上接载乘客，并且可以在任何地方落客。每辆Green TAXI出租车也都可以在曼哈顿北部，布朗克斯，皇后区，布鲁克林和史坦顿岛和机场进行预约载客。我们先来看下1-6月的整体使用趋势。

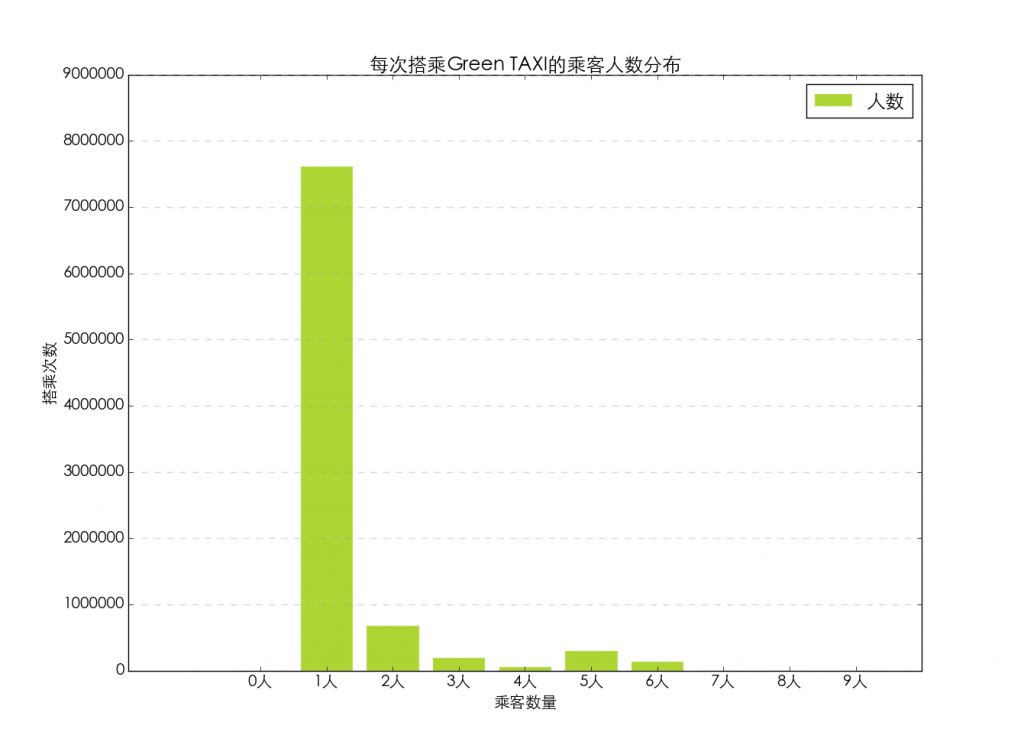
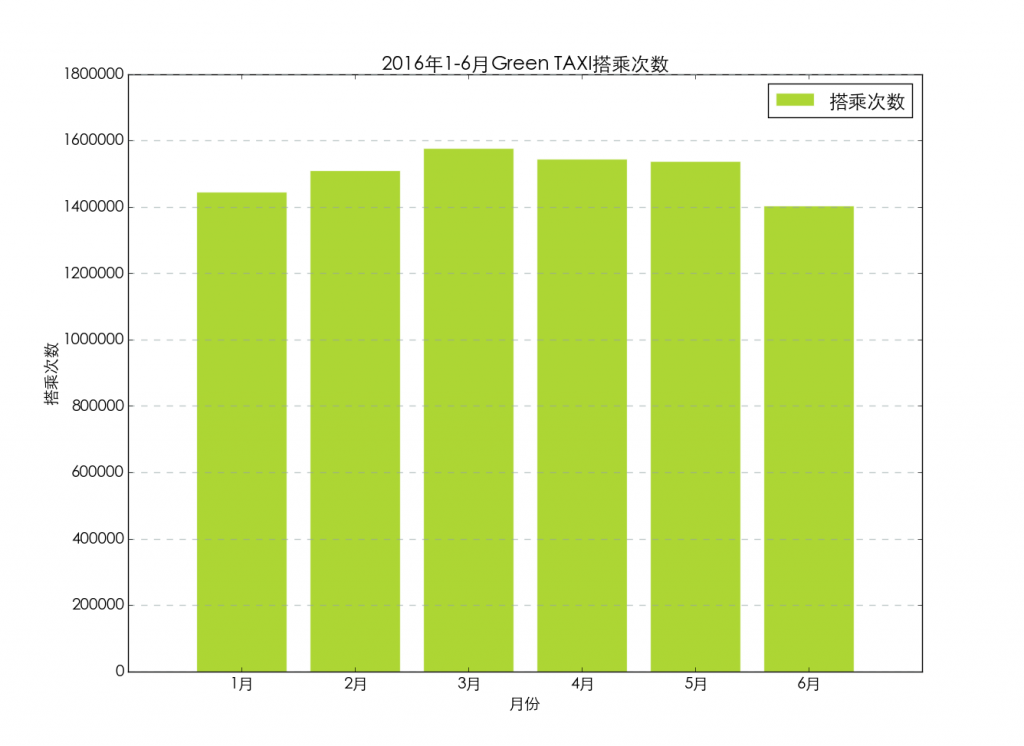
[](http://bluewhale.cc/wp-content/uploads/2017/01/map.jpg)

我们对Green TAXI中的数据进行处理，获得按月的使用量变化趋势。首先将数据表中的载客时间字段转化为日期格式，然后把这个字段设置为数据表的索引字段。并按月的维度对数据表中的数据进行汇总计数。并提取VendorID列作为每月Green TAXI的载客数量。

将按月的Green TAXI数据绘制为柱状图，来查看整体的变化趋势情况，下面是绘制趋势图的结果。

  
  
 从图表来看，1月2月和6月的使用量相对较低，3月到5月使用量相对较高。这样的趋势与季节和温度有关系吗？纽约冬天一般从11月到次年3月，冷且风大。出租车的使用量是否和气温有关呢？我们这里先做一个假设，在文章的最后再来分析。

## 2、乘客数量分布

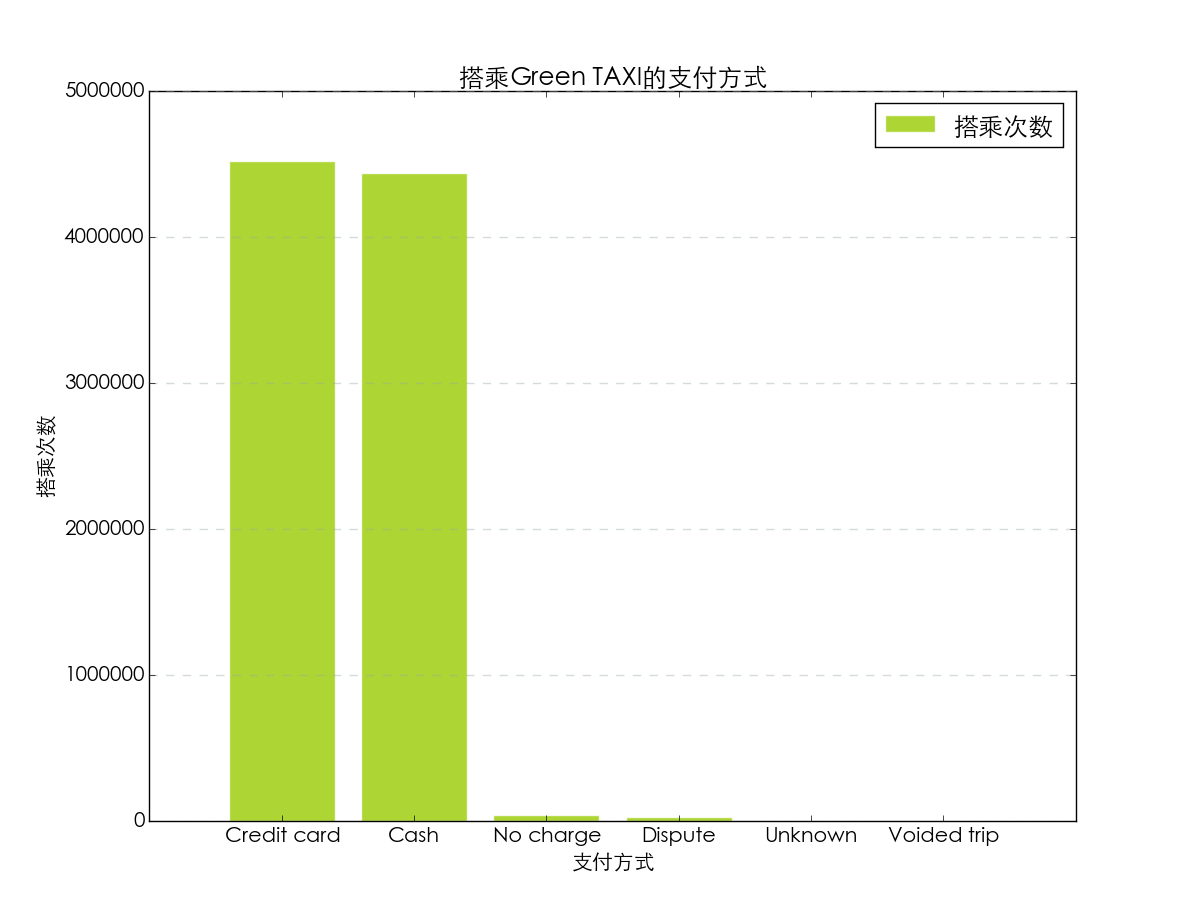
Green TAXI 的车型分为四种，根据用途和载客数量不同从小到大分布为Sedan，Town car，Minivan和SUV。下面我们来看下搭乘Green TAXI的乘客数量分布。换句话说就是每次搭乘时出租车内的人数。这里需要说明的是这个数字的采集方式是由出租车驾驶员手动输入的。因此可能会有一些不准确性。  
[](http://bluewhale.cc/wp-content/uploads/2017/01/eltaxistanews_taxi-en-edificio-de-TLC.jpg)  
 要获得每次搭乘出租车乘客的分布情况，需要对数据表进行处理。我们首先查看下搭乘人数的范围，查看Passenger\_count的最大值和最小值，范围是0-9个人。这里就可以看出问题。首先0人搭乘这个明显是有问题的。莫非是约车的时候打表来接的，然后又取消了订单？其次9个人乘坐出租车这个应该是SUV车型，应该是包含儿童或未成年人。否则即使是7座的SUV也很难坐下9个成年人。  
  
  
  
  
 从乘客数量的分布情况来看，独自1人搭乘出租车的数量最多，其次为2人，5人，3人和6人。这里也有可能是出租车司机输入时的固定选择造成的。现有的数据中没有可以进行交叉验证的数据，因此这个乘客人数分布数据仅供参考。

## 3、支付方式

Green TAXI数据表中对乘客支付车费的方式分为了6类，通过对应文档分别为Credit card,Cash,No charge,Dispute,Unknown,Voided trip。但仔细看会发现，这并不是6种不同的支付方式。如后面的免费，争议和空驶。甚至还出现了Unknown的情况。不过我们还是按Payment\_type字段对数据进行了汇总统计。下面是支付方式的结果。

[](http://bluewhale.cc/wp-content/uploads/2017/01/payleven_taxi.jpg)

首先将支付方式的编号还原为具体的类别名称。然后按支付方式的名称字段对数据进行汇总。再对汇总后的支付方式数据汇总成图表进行分析。



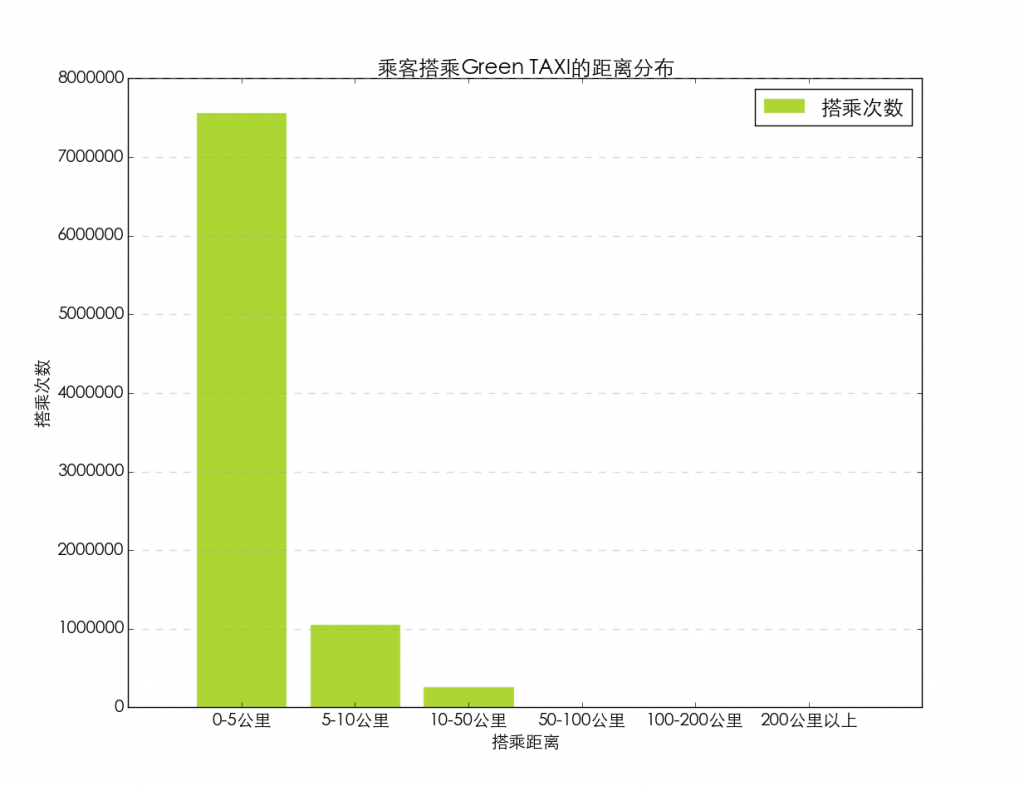
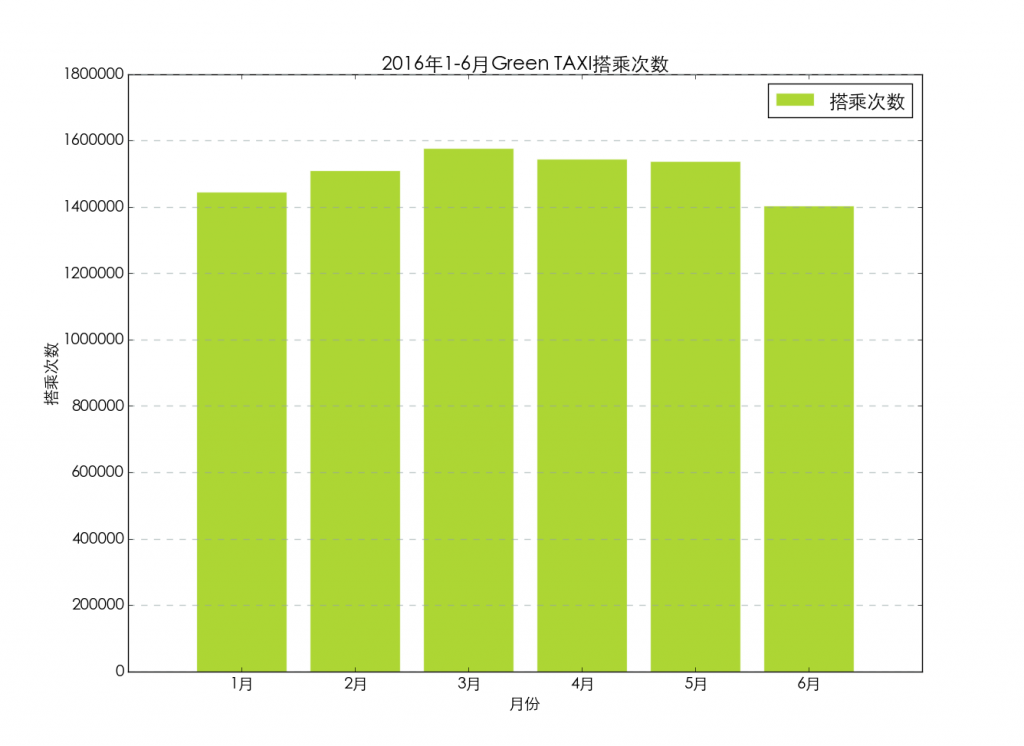
在所有的支付方式中，现金和信用卡是乘客使用最多的支付方式。其他的情况都很少见。其实本来几个类别就不是支付方式的分类。

## 4、平均距离及里程分布

Green TAXI起步价2.5美元（0.2英里以内），之后每0.2英里（约320米）或者等候2分钟加收40美分。从晚间8点到早上6点期间，加收夜行附加费0.50美元。高峰时刻（周一到周五下午4点到8点）附加费1美元。此外，乘客还需承担乘车期间产生的任何费用并另付小费（15%以上）。这里搭乘距离是影响金额的主要因素，我们来看下乘客在搭乘出租车时的距离分布情况。



首先来看下乘客搭乘出租车的距离范围，查看Trip\_distance的最大值和最小值。最短距离为0，最大距离为832.2英里。这里的0英里不知道是什么情况。

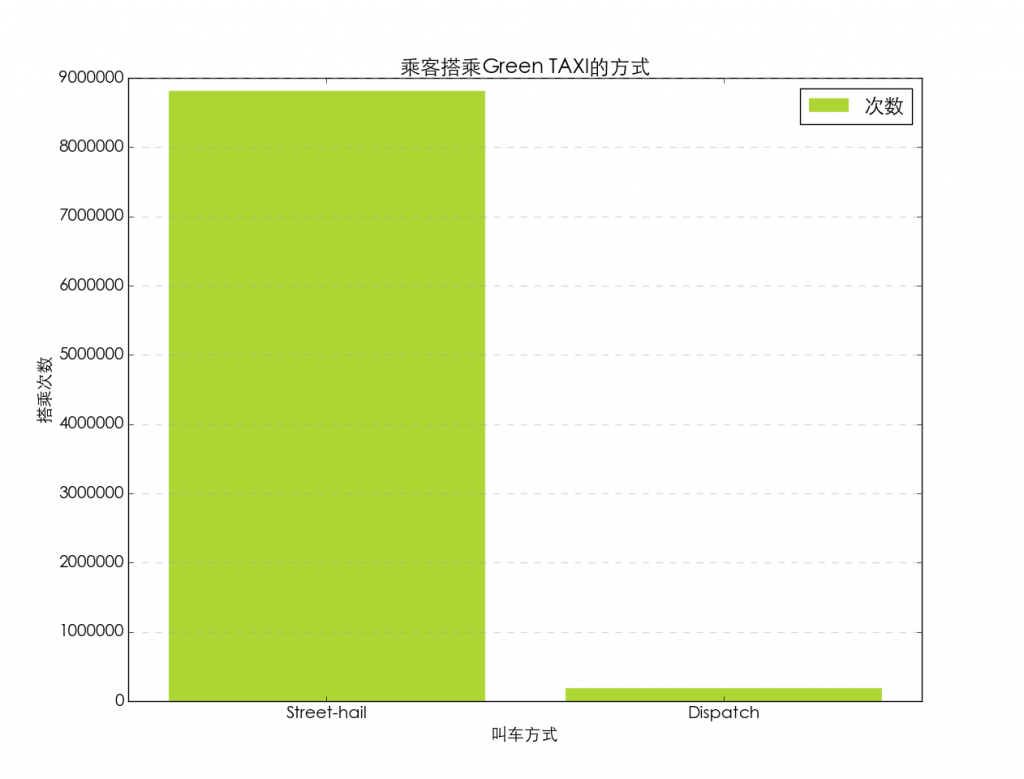
首先来看下乘客搭乘出租车的距离范围，查看Trip\_distance的最大值和最小值。最短距离为0，最大距离为832.2英里。用总的搭乘里程除以乘客搭乘次数计算出每次搭乘的平均行驶距离为2.81英里。对乘客搭乘距离进行分组，以5公里为一组进行划分。  
  
 按划分后的距离分组字段对数据表进行计数汇总，查看乘客搭乘的距离分布情况。  
将乘客搭乘的距离分布数据绘制成图表  
  
  
  
从乘客搭乘距离分布图上来看，0-5公里短途的数量最多，随着距离的增加搭乘次数明显减少。很明显长途出行搭乘出租车不是一个经济的选择

## 5、乘客叫车方式

Green TAXI的定位是Yellow TAXI的补充，因此对于Green TAXI的运营区域也有严格的限制，再来看下Green TAXI的载客区域。Green TAXI可以从曼哈顿北部（西110街和东96街的北部），布朗克斯，皇后区（不包括机场），布鲁克林和史坦顿岛的街道上接载乘客，并且可以在任何地方落客。但在曼哈顿北部，布朗克斯，皇后区，布鲁克林和史坦顿岛和机场这些区域，Green TAXI只能进行预约载客。也就是说，如果乘客没有预约，Green TAXI就不能去这些区域载客。下面我们来看下Green TAXI的数据中乘客叫车的方式。

[](http://bluewhale.cc/wp-content/uploads/2017/01/119456484.jpg)

数据表中的Trip\_type字段代表了乘客召唤出租车的方式，1= Street-hail，2= Dispatch。我们按Trip\_type字段对数据进行汇总。



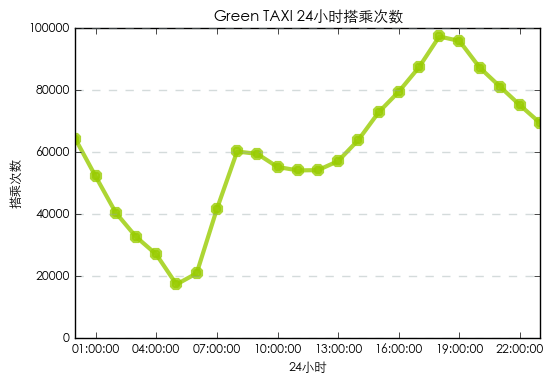
从图表中看到，Street-hail即乘客在路边召唤出租车的方式是最主要的叫车方式。Dispatch这类通过调度方式叫车的数量非常少。这里有一点需要说明，2016年1月-6月除了Green TAXI和Yellow TAXI外，还有Uber存在。因此乘客在预约时可能更多的选择了Uber，而非传统的出租车。这里没有Uber同期的数据，仅仅是猜测。

## 6、24小时使用趋势

我们再来看下乘客在一天24小时中使用Green TAXI的情况。这里我们只使用了2016年1月的单月数据进行分析。

[](http://bluewhale.cc/wp-content/uploads/2017/01/1.jpg)

重新导入1月的数据，并提取出每次搭乘的小时数据，将数据表按小时数据进行汇总。绘制24小时Green TAXI的使用变化趋势图。

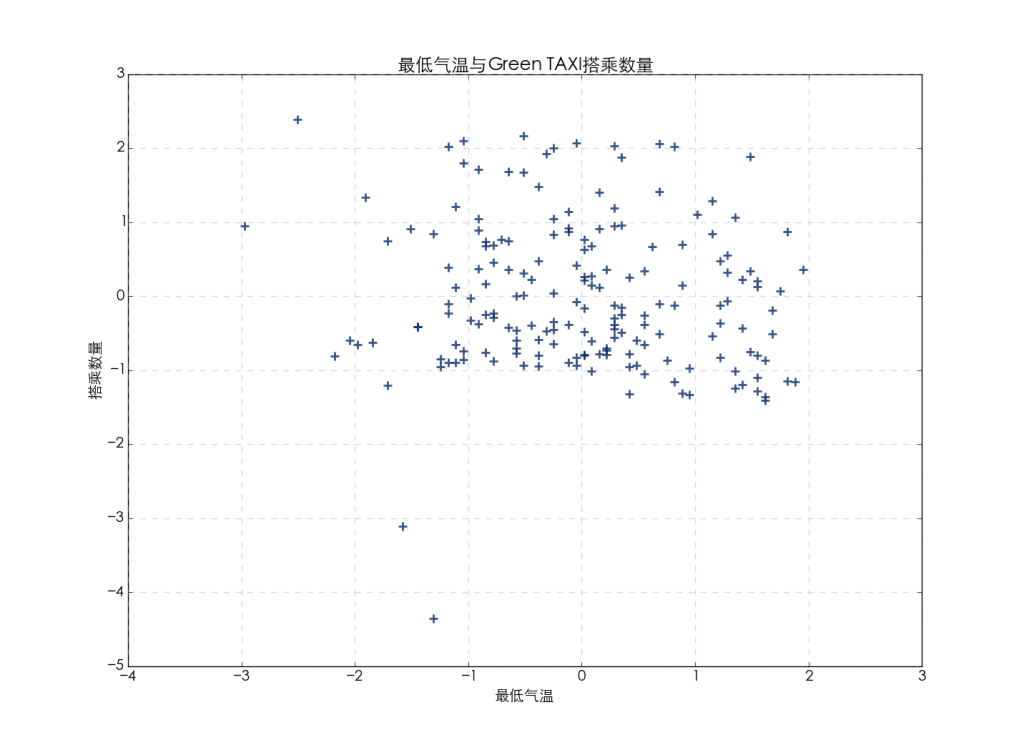
前面完成了24小时数据搭乘数据的汇总，下面绘制24小时搭乘趋势图。   
  


从24小时搭乘趋势图来看，每日晚间是Green TAXI的使用高峰。晚间的出租车使用量明显高于早间，并且延续时间较长，从晚间19点的用车高峰一直持续到凌晨。

## 7、气温与出租车使用情况的关系

本文开始时，我们在分析1-6月Green TAXI使用趋势时曾假设出租车的使用与气温有关，冬季由于气温较低造成出租车的使用量也较低。下面我们使用1月-6月的天气数据与出租车的搭乘数据进行相关分析，来看下这两者间是否有关联。



首先对出租车搭乘数据进行预处理，包括更改字段格式以及对数据进行按天汇总和提取工资。导入1-6月的天气数据，然后从中提取最低气温数据，也就是TMIN字段中的值。用于和出租车搭乘数据进行相关分析。  
  
将预处理后的数据绘制散点图，下面是具体图表。  
  


从散点图中来看，最低气温与出租车的搭乘数量间并没有明显的联系。出租车的搭乘数量一致维持固定的区间内，并没有因为最低气温的变化有显著的变化。下面我们使用回归分析计算两者的相关性。

将每日的最低气温设置为自变量X，每日出租车的载客量设置为因变量Y。然后使用线性回归模型来计算判定系数R方。

判断系数为0.0009，说明自变量对因变量的解释度较低，换句话说最低气温的变化与出租车载客量间没有联系。

# 八、项目分析步骤

1、源数据和业务需求分析

根据数据仓库自上而下和自下而上的建模思想需要做一下分析：

根据数据字典和源数据分析源数据中的每一项指标和属性；

分析业务需求；

**纽约出租车数据字典**

|  |
| --- |
| 数据字段描述：  medallion：UUID  hack\_license：UUID  vendor\_id：类型  rate\_code：比率  store\_and\_fwd\_flag:是否是四驱  pickup\_datatime：客人上车时间  dropoff\_datatime：客人下车时间  passenger\_count：载客数量  trip\_time\_in\_secs：载客时间  trip\_distance：载客距离  pickup\_longitude：客人上车经度  pickup\_latitude：客人上车维度  dropoff\_longitude：客人下车经度  dropoff\_latitude：客人下车维度 |

**源样例数据如下图：**

|  |
| --- |
|  |

## 2、逻辑模型建立

根据对源数据和业务需求的分析情况，建立数据仓库的基础层、汇总层和应用层逻辑模型。

逻辑模型就是确定我们需要的数据指标、字段；分析的维度、度量、粒度等信息以及维度、度量之间的关系。

## 3、大数据平台建立

Hadoop大数据平台建立包括Hdfs、Mapreduce和Hive等。

**大数据平台数据目录结构如下图**：

|  |
| --- |
|  |

## 4、物理模型创建

在Hive的大数据仓库上建立数据仓库的物理模型。

**创建脚步如下所示：**

|  |
| --- |
| create external table trip(  medallion string,  hack\_license string,  vendor\_id string,  rate\_code string,  store\_and\_fwd\_flag string,  pickup\_datatime string,  dropoff\_datatime string,  passenger\_count bigint,  trip\_time\_in\_secs bigint,  trip\_distance int,  pickup\_longitude string,  pickup\_latitude string,  dropoff\_longitude string,  dropoff\_latitude string  )  ROW FORMAT DELIMITED  FIELDS TERMINATED BY '|'  LINES TERMINATED BY '\n'  STORED AS TEXTFILE LOCATION '/trip\_input/'; |

## 5、ETL实现

通过Flume采集数据到ODS层；

通过Mapreduce来转码ODS层数据到数据仓库的基础层；

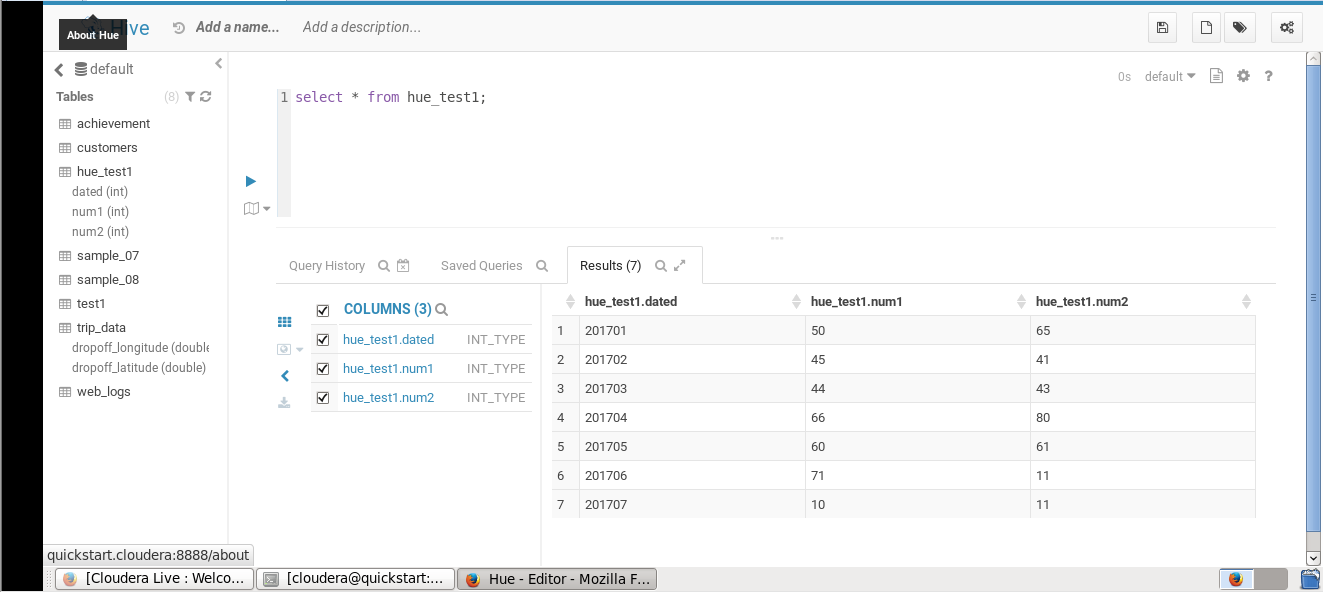
通过HiveSQL汇总数据到数据仓库的汇总层；

通过HiveSQL生产应用层数据；

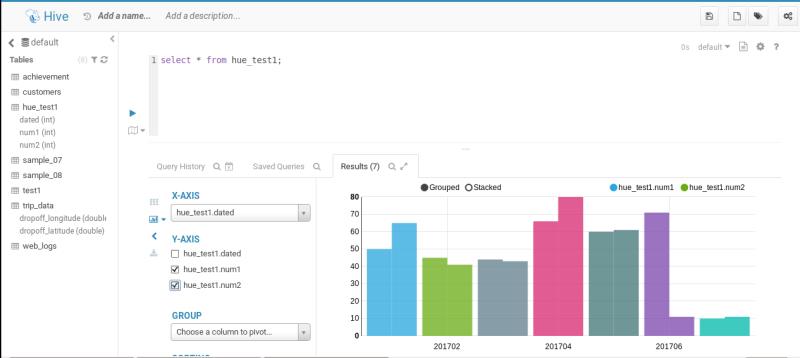
1. 将数据集下载到linux本机
2. 在hive下建立数据仓库
3. 将数据导入hive仓库
4. 通过mapreduce提取需要的计算，进行处理计算
5. 将计算结果放入一个hive表里
6. 将hive表内容导入hue，绘制最后的图形结果

## 6、通过Hue图形化进行数据的分析和展现

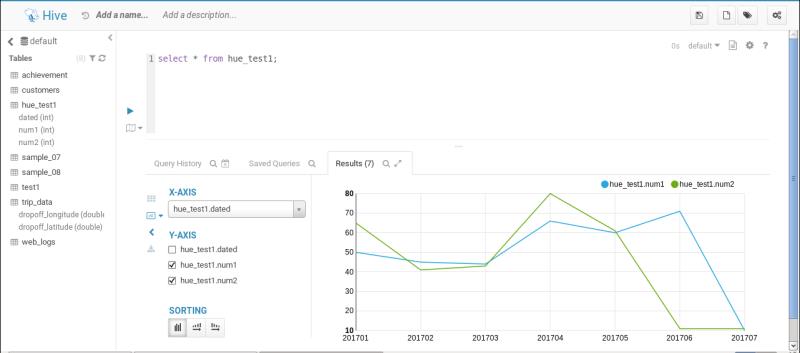
如下图1~3所示：



（图1）

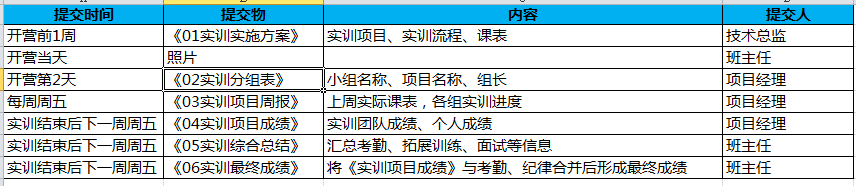


（图2）



（图3）

# 九、项目提交物



# 十二、实训标准

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **实训 类型** | **实训 规划** | **实训 内容** | **达标 要求** | **实训 目的** |
| 基础类 实训 | 岗前制度培训 | 礼仪、考勤、日报、周报制度等 | 着工装准时到岗、准时提交日报与周报 | 建立员工基本素养与协调沟通能力 |
| 体验实训 | 安装部署项目，按用户手册使用系统 | 能使用系统完成业务操作 | 建立业务背景 |
| 技术结构分析与总结 | 分析系统使用的技术结构 | 提交技术结构体系图与总结 | 总结所学技术 |
| 阅读分析设计文档 | 在老师指导下阅读项目的分析设计文档 | 熟悉编码步骤，提交编码计划 | 熟悉项目业务逻辑 |
| 熟悉任务模块与设计 | 明确各自任务模块、搭建项目工程环境 | 进入编码前准备状态 |
| 编码与业务逻辑实现 | 编码，特殊技术攻关 | 提交模块源代码 | 累积编码经验、自主技术学习能力 |
| 集成与测试 | 集成模块到系统，并测试 | 系统能运行并通过测试。提交该模块用户使用手册 | 累积集成与测试经验 |
| 项目总结 | 答辩项目实现过程与创新 | 提交PPT，答辩评分通过 | 培养答辩，简报制作能力 |

# 十三、实训学员日常管理制度

在学生的日常管理方面，严格按照企业员工的管理规范进行，采用科学、多样的管理方法，力求还原企业的实际工作环境，营造出紧张有序的工作氛围，以期达到学生实习的真正目的。

**日报制**

日报制度，是现代软件企业管理员工的一个通用方法，也是新员工入职后的第一门必修课，主要是员工对每天个人工作内容的总结，以及记录对公司文化、技术情况的认知和理解。对于实习的学生，将采取日报制度，使学生明了自己的工作职责和工作重点、清楚自己当天的工作内容，并及时对自己的工作内容进行监督和评估。

**项目组制度**

按照现代软件企业通用的软件开发岗位的工作模式，从实习开始，根据学生的情况进行项目组的划分，在整个实习过程中以项目组为单位组织活动和学习。每组有一名组长，相当于项目经理，带领项目组完成任务，并就组内成员的技术问题展开讨论；老师以CTO（技术总监）的身份教学并指导项目组的工作。项目组的划分将使学员一开始就会产生企业工作的角色感，并在整个过程中不断提高个人的团队合作和职业意识。

**例会制度**

实习期间，按照相关制度要求，安排例会，分为组内例会和负责人例会，例会制度增进学生们的沟通，并能及时根据沟通情况反馈学生学习状况。

**调查问卷**

将定期组织问卷调查，针对教师及教学实施方面存在的情况进行反馈，及时纠正教学过程中出现的问题，保证实习的最终效果。

# 十四、实训周报模板

**实训项目周报**

**院校：**XXXX大学 **周次：**第1周

实际实训安排

|  |  |  |
| --- | --- | --- |
| **星期** | **上午** | **下午** |
| 1 | 开营仪式 | 项目分组，需求说明 |
| 2 | 讲课：Web UI设计 | 实战：需求分析 |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |

各组进度

|  |  |
| --- | --- |
| **小组名称** | **进度** |
| 组名1 |  |
|  |  |
|  |  |
|  |  |
|  |  |

其他事项

此处列出本周发生的与项目实训有关的事宜

# 十五、实训考核制度及体系

## 1、成绩得分

总成绩 = 团队得分\*40% + 项目组长评分40% + 项目经理评分\*20%

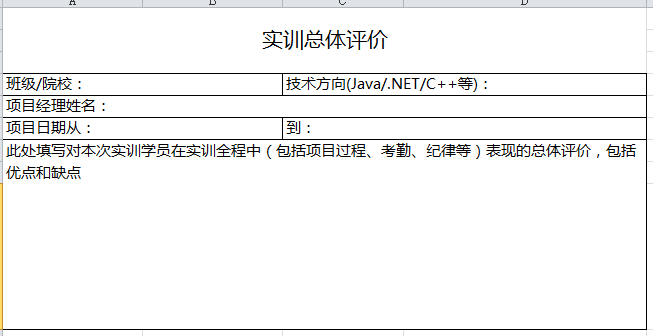
团队得分 = 原型设计\*20% + 系统设计\*20% + 项目提交\*60%

项目组长评分：由各小组项目组长根据组员表现给出

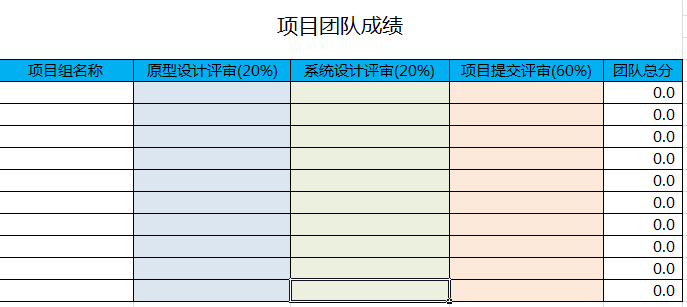
项目经理评分：由项目经理根据对学员的代码审查结果及日常表现给出

学生不准迟到，缺课

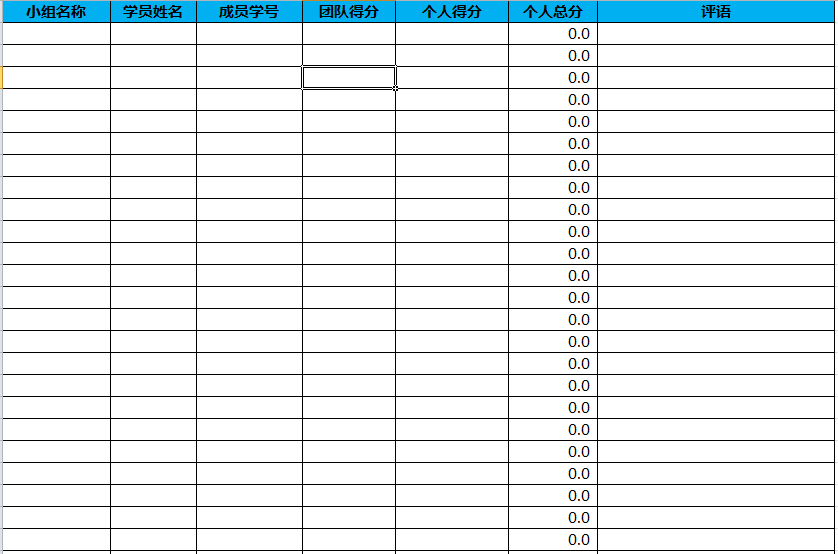
## 2、项目结束评价



## 3、团队项目成绩



## 4、个人成绩



## 5、实训最终成绩

