**数据预处理✅**

**去除订单开始/结束日期、地点为空及非12月内的记录➡️**yellow/green/fhv\_tripdata\_2024-12\_cleaned.parquet

**基础描述统计✅**

descriptive\_analysis.py

basic\_visualization.py

**Yellow Taxi (黄色出租车)**

作为纽约市最主要的出租车服务，黄色出租车的分析结果最能反映出城市的核心出行脉络。

* **时间节律：**
  + **每日需求：** 黄色出租车的需求高峰非常集中，**在下午4点到晚上7点之间**达到顶峰，这段时间的需求量超过了全天总订单数的1/5。这完美契合了城市居民的下班通勤和晚餐、夜间活动出行。从凌晨开始，订单量缓慢上升，在上午10点后开始加速，并在下午形成稳定的高需求平台，直到深夜才逐渐回落。
  + **每周需求：** 黄色出租车的周需求量非常均衡，表明其服务于日常的、普遍的出行需求。周二和周一的需求量略高于其他日子，而周三则相对较低，这可能与特定周内的活动安排有关。
  + **每日高低峰：** **12月12日**是该月订单量最高的一天，达到了143,442单，而\*\*12月25日（圣诞节）\*\*则以47,766单成为最低谷，这强有力地证明了节日对公共交通出行的显著影响。
* **服务画像：**
  + **行程特征：** 平均行程距离为 **3.34英里**，平均时速仅为 **10.53英里/小时**。这两个数据共同描绘了其在城市中心拥堵区域缓慢前行的典型场景。
  + **成本与效率：** 平均总费用为 **29.49美元**，但平均单位距离成本高达 **20.69美元/英里**。这反映了其服务模式下，计费不仅基于距离，还包含了因交通拥堵导致的低速等待时间。
  + **支付方式：** \*\*超过80%\*\*的乘客使用信用卡支付（81.79%），现金支付仅占14.68%，这表明乘客更倾向于使用便捷的电子支付方式。
* **异常值：** 值得注意的是，黄色出租车有大量异常值记录，包括 **44万条超长距离**、**19万条超长时间**和 **35万条超高费用**的记录。这表明在数据收集或录入过程中可能存在错误，或者这些异常值反映了非常规的长途高价行程。

**Green Taxi (绿色出租车)**

绿色出租车主要服务于纽约市的非核心区域，其数据特征与黄色出租车有明显差异。

* **时间节律：**
  + **每日需求：** 与黄色出租车相似，其需求高峰也在**傍晚5点到7点**，但订单量远小于黄色出租车。这表明它可能服务于非核心区域的下班通勤。
  + **每周需求：** 需求量同样在工作日相对稳定，**周二和周一**需求量最高，周末需求量有所下降。这符合其服务区域的居民通勤特点。
  + **每日高低峰：** 同样在**12月12日**达到订单峰值（2,370单），并在**12月25日**降至低谷（773单），这与黄色出租车的趋势完全一致。
* **服务画像：**
  + **行程特征：** **平均行程距离高达14.32英里**，远超黄色出租车。更令人惊讶的是，**平均时速达到了61.54英里/小时**。这个速度非常不寻常，强烈暗示绿色出租车承担了大量往返于机场或郊区的高速公路行程。这个极高的时速需要进一步的数据验证。
  + **成本与效率：** 单位距离成本为 **18.98美元/英里**，低于黄色出租车。这可能是因为长途行驶在高速公路上，减少了低速等待时间，提高了效率，从而降低了单位成本。
  + **支付方式：** 绿色出租车的信用卡支付比例为 **73.03%**，低于黄色出租车，而现金支付比例为 **26.02%**，高于黄色出租车。

**FHV (电召车)**

FHV的数据存在缺失，但已有的信息足以揭示其独特的运营模式。

* **时间节律：**
  + **每日需求：** FHV的每日高峰与两种出租车都不同，集中在**上午9点到下午1点**，这是典型的商务出行或办事高峰期。其在晚高峰时段的需求量反而较低。
  + **每周需求：** 需求量在周一和周二最高，而**周末（周六和周日）的需求量显著下降**，这进一步印证了其服务于工作日通勤和有计划的出行。
* **服务画像：**
  + **数据缺失：** 遗憾的是，由于缺少 trip\_distance 和 payment\_type 等关键字段，无法进行更深入的服务画像分析。我们只能从 **10,060条超长时间异常记录**中推断，FHV可能存在一些非常规的长途行程。

**结论**

综合来看，这三种服务模式满足了不同的城市出行需求：

* **黄色出租车**：是应对城市核心区域日常、高密度、短途出行需求的主力军。
* **绿色出租车**：更像是一个补充者，主要服务于城市周边、郊区以及需要走高速公路的长途出行。
* **FHV**：则更专注于工作日，服务于有计划的、非晚高峰的商务或办事出行。

**Yellow✅/Green✅/FHV✅连地理信息&可视化**

fclass.py➡️all\_fclass\_types.txt手动分类➡️merging\_position\_pickup.py➡️fused\_pickup\_yellow.parquet➡️POI\_visualization.py➡️poi\_pickup\_map\_yellow.html

merging\_position\_dropoff同理

不同服务类型的上下车点POI类型占比差异（可以做个折线图可视化，不急）：

**核心发现：FHV 与 Yellow 的高度相似性**

通过对比 FHV 和 Yellow 的数据，我们可以发现它们的出行模式在POI类型分布上达到了惊人的相似程度。

* **共同的核心驱动力：** 两种服务都以**基础设施**和**食品**作为最主要的出行类型。在FHV的上车点中，基础设施占比为**41.35%**，食品为**23.37%**。同样，在Yellow的下车点中，基础设施占比为**41.39%**，食品为**23.32%**。这两大类别合计贡献了超过 64% 的出行量，表明它们是城市日常运转和居民基础生活的核心需求。
* **均衡的次要类型：** 在其他如**购物**、**生活服务**和**文娱**等类别中，两者的占比也几乎完全一致。这种高度的同质性说明，FHV和Yellow作为主要服务于城市核心区的交通工具，其服务功能和用户群体高度重叠。

**绿车：服务模式的微妙差异**

相较于FHV和Yellow，Green Taxi的数据虽然大体相似，但在一些细节上呈现出其独特的服务模式。

* **相似的基础结构：** 像其他两种服务一样，Green Taxi的**基础设施**（41.51%）和**食品**（23.35%）也占据了上车点的前两位。这表明其服务区域的居民同样以这两类POI为主要出行起点。
* **潜在的区域差异：** 值得注意的是，Green Taxi在**生活服务**上的上车点占比为**9.15%**，略高于Yellow的**9.03%**。这可能暗示Green Taxi服务的非核心区域，其居民对这类日常服务场所（如理发店、洗衣房等）的依赖性稍强。

**内部对比：上车点与下车点的平衡性**

通过对比每种服务内部的上车点和下车点，我们可以更深入地洞察乘客的出行流向。

* **FHV 和 Yellow：** 这两种服务的上车和下车POI占比几乎完全对称。例如，FHV在**旅游景点**的上车占比为**2.98%**，下车占比为**3.01%**。Yellow在**教育**上的上车占比为**1.58%**，下车占比为**1.58%**。这种高度的平衡性表明，这两种服务主要满足的是双向、往返式的城市出行需求。
* **Green Taxi：** 尽管总体均衡，但Green Taxi的上车和下车点存在一些细微的、有趣的差异。
  + **下车点略高：** 在**食品**、**购物**和**文娱**这三类POI中，下车点的占比都略高于上车点（例如，**食品**：下车23.71% vs 上车23.35%）。
  + **上车点略高：** 相反，在**基础设施**和**生活服务**中，上车点占比则略高于下车点。 这些趋势表明，乘客使用Green Taxi时，更倾向于以**食品、购物、文娱**为目的地，而以**基础设施、生活服务**为出发点。这可能反映了其服务区域的用户出行模式，比如从家中或办公地点（基础设施）前往购物或娱乐场所。

**Yellow/Green/FHV连天气信息&可视化✅**

merging\_weather.py➡️三份融合好的文件

weather\_temp\_visibility.py➡️不同天气/能见度/气温下订单数 统计\*3

差异性检验：天气🈶能见度🈚️

Efficiency\_cost.py➡️黄&绿 天气对效率（平均行程长度和时长）和成本（总费用和单位价格）的影响及差异性检验

Peakhours\_week.py➡️分早晚高峰 周末工作日差异性检验坏天气对订单量是否有影响\*3

可视化——————————————————————————————————

max\_min\_path\_visibility.py➡️12/12&12/25的yellow&green的单日乘车路径可视化（结合版：multi\_map\_dashboard.html；四个小html：map\_green\_25、map\_green\_12、map\_yellow\_25、map\_yellow\_12）

**异常值（时间、价格、距离）分析 特征工程 天气/时间/……的影响✅**

三个outliers\_records.parquet（描述统计得出）

把异常值的数据文件with天气、地理（特征是“是否是雨雪天气”、“是否是周末”、“如果是工作日是否是早晚高峰”、“上车点和下车点的POI类别”）  
——>逻辑回归。结果是：

1. Yellow的异常订单均为POI不明确的情况下产生
2. Green和FHV是正常的

**对于Yellow&Green进一步：**

**乘客：已知时间、天气、出发地，选哪种服务➡️预测价格&时间**

因为数据量限制➡️only green

先把大多数据放入spss进行降维➡️得五个因子：

因子1 (降水因子)

该成分在 precip（降水量，0.280）、precipprob（降水概率，0.307）和 is\_rain\_snow（是否雨雪，0.288）上显示出强正向载荷。这些变量都与降水的发生和可能性密切相关。它们之间的高度正相关性表明，该因子能够综合反映整体的降水事件，因为在潮湿天气条件下，这些指标往往同步上升。

因子2 (体感舒适度因子)

此因子的最高载荷体现在 feelslike（体感温度，0.315）、temp（温度，0.314）和 dew（露点，0.245）上，同时在 sealevelpressure（海平面气压，-0.213）上呈负向载荷。这些变量均与温度感知和湿度紧密相连，共同影响着人们的体感舒适度。例如，露点与温度结合，能够决定体感冷热，因此这个因子可以被看作是与舒适度相关的变量集合。

因子3 (太阳辐射因子)

该因子由 solarradiation（太阳辐射，0.338）、solarenergy（太阳能，0.338）和 uvindex（紫外线指数，0.334）主导。这些变量都是阳光强度的直接测量指标，因此该因子很自然地代表了太阳活动，它通常独立于风力或降水等其他气象要素而变化。

因子4 (风力动态因子)

主要载荷变量包括 windgust（阵风，0.397）、windspeed（风速，0.390）、winddir（风向，0.182）以及 sealevelpressure（海平面气压，-0.229）。风速和阵风是大气运动的核心，而气压往往会影响风的模式（例如低气压可能伴随强风）。该因子似乎反映了风的行为及其相关动态。

因子5 (积雪因子)

此因子的最强载荷来自 snow（降雪量，0.723）和 snowdepth（积雪深度，0.504），并在 is\_rain\_snow（是否雨雪，0.119）上有适中载荷。这些变量专门追踪降雪量和积雪的累积情况，将其与一般的降水（如因子1所示）区分开来。这符合逻辑，因为降雪是一种独特的冬季气象现象。

用到了前面分析的（是否为工作日、是否为早晚高峰）

地理数据：先加入了上车点和下车点的POI类型，并加入出租车数据中自带的trip\_distances

出租车数据中自带的：乘客数量、支付方式、支付手续费

————————

以上所有数据用于预测在车上的时间&价格，结论（符合现实状况）：

预测时间：R2在0.65~0.7

预测价格：R2近0.8

**司机：已知服务类型、时间、天气➡️接客地点**