交通大数据 | 2200 万次纽约公共自行车骑行的故事：纽约公共自行车系统大数据分析

一、纽约公共自行车系统的一天

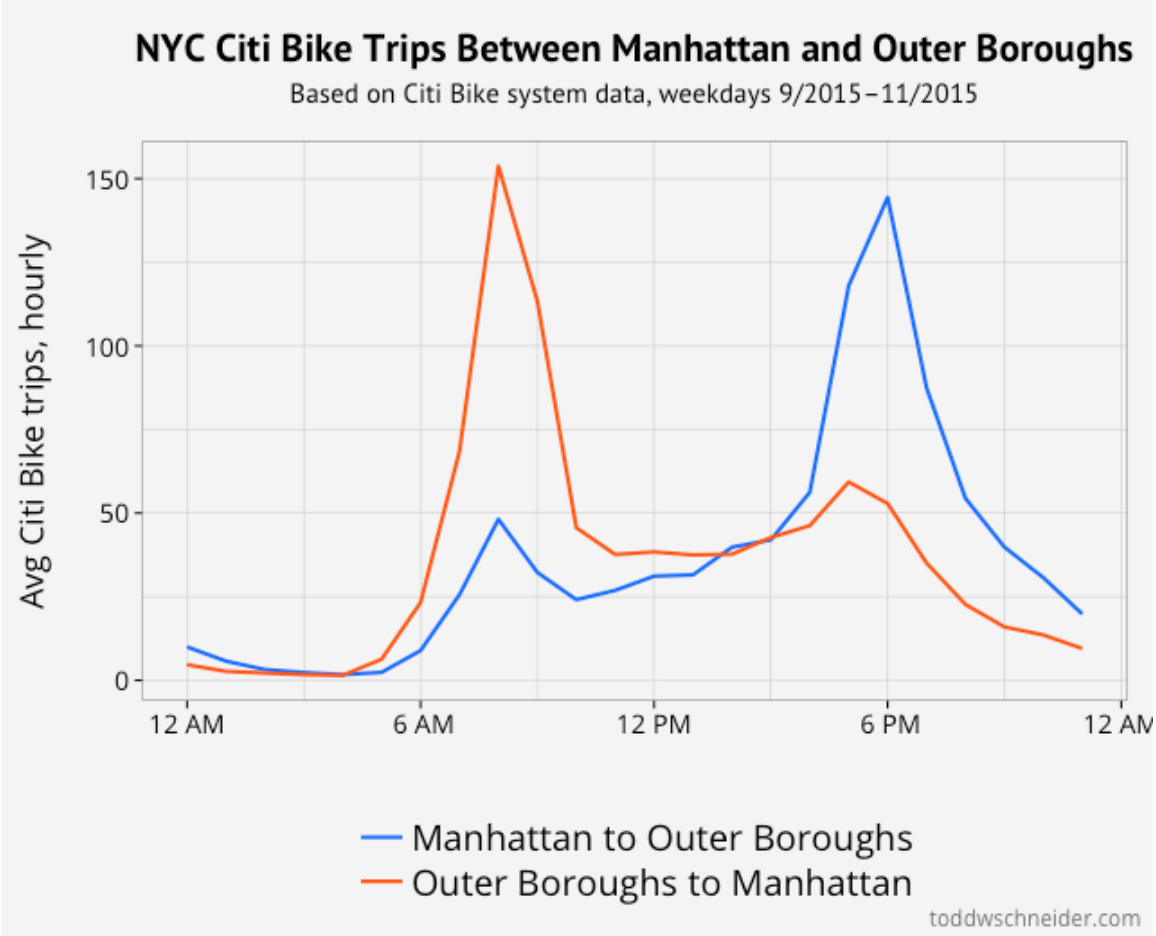
我获得了纽约公共自行车系统 2015 年 9 月 16 日星期三的出行数据，假设每一次出行遵照谷歌地图推荐的骑行导航，利用 CartoDB 的开源资料库 Torque.js 制作了一个动画。那天一共有 51179 条出行记录，但是我排除了开始和结束于同一站点的记录，在可视化中剩下 47969 条出行记录。地图上的每一个蓝色点表示一次单程的公共自行车出行，小的橘色的点表示分散在城市内的 493 个公共自行车站点：

图一、动画

如果你盯着动画看一会，你开始看到一些趋势。我自己最喜欢的观察地点是连接布鲁克林区和下曼哈顿区的桥。早上 8 点左右开始，你看到稳定的自行车流通过布鲁克林桥、曼哈

顿桥和威廉斯堡桥从布鲁克林到曼哈顿。中午这些桥一般都不太忙了，然后从下午 5 点 30 分左右开始，我们看到蓝色的点从曼哈顿流回到布鲁克林，是因为骑行者们离开他们在曼哈顿的办公室返回他们在布鲁克林的家。

通过观察曼哈顿和外部区域之间的每小时出行量图，我们可以直接从数据中观察到这一现象：



图二、曼哈顿和外部区域之间的纽约公共自行车出行量（基于 2015 年 9 月-11 月间工作日的纽约公共自行车数据）

果然，早上从布鲁克林到曼哈顿的骑行量要多于相反方向，而晚上有更多的人从曼哈顿骑车到布鲁克林。很值得注意的是，绝大部分的公共自行车出行开始并结束于曼哈顿。自从 2015 年 8 月项目扩容以来，整个系统分列了：

88% 的出行量开始并结束于曼哈顿

8%的出行量开始并结束于一个外部区

4%的出行量在曼哈顿和一个外部区之间骑行

在动画中还有其他明显的通勤模式：早上从 59 街向北延伸到第一大道的公共自行车流量很少，但是下午 5 点左右开始流量回升，大概是因为人们从他们位于中城的办公室回到上东区的家。

同样地，如果我们看早高峰从下东区穿过默里山延伸至第一大道和第二大道的平行路段，显然这里有更多的沿着第一大道向北进入中城的流量。在晚上有更多的沿着第二大道向南的流量，因为上班族们回到住宅区。

如果我们利用 2015 年 8 月纽约公共自行车系统扩容以来的所有出行记录，再一次假设每一次出行遵照谷歌地图推荐的导航，我们可以看到整个城市哪些路段是公共自行车经过最多的。这里有一张展示最受欢迎的道路的地图，图中线条的厚度和亮度是基于这一路段经过的公共自行车数量：


NYC Citi Bike Most Popular Roads
Sep—Nov 2015
toddwschneider.com

图三、纽约公共自行车最受欢迎的道路

这张地图让人想起我以前的文章中的出租车上、下客地图，但是它们实际上略有不同。出租车的地图是由单个的点组成，每个点是一次上客或下客，而上面的纽约公共自行车地图把每一次出行看成是从起点到终点的一连串的线段。

这张地图为骑行者们展示了为数不多的主要路线：在西边第八大道和第九大道分别朝向上城区和下城区，在东边第一大道和第二大道分别朝向上城区和下城区。公共自行车经过最多的一条路段位于第八大道从 28 街到 29 街的路段。其他主要的自行车路线包括百老汇、斜线穿过曼哈顿中城和西区沿着哈德逊河的自行车道。

记住地图和动画假设人们遵照谷歌地图的导航，这绝对不总是对的。谷歌地图似乎对有专用自行车道的道路表现出强烈的偏好，这就是为什么，比如第八大道有很多流向上城的交通量，但第六大道很少。两条大道都是向北的，但只有第八大道有一条专用自行车道。

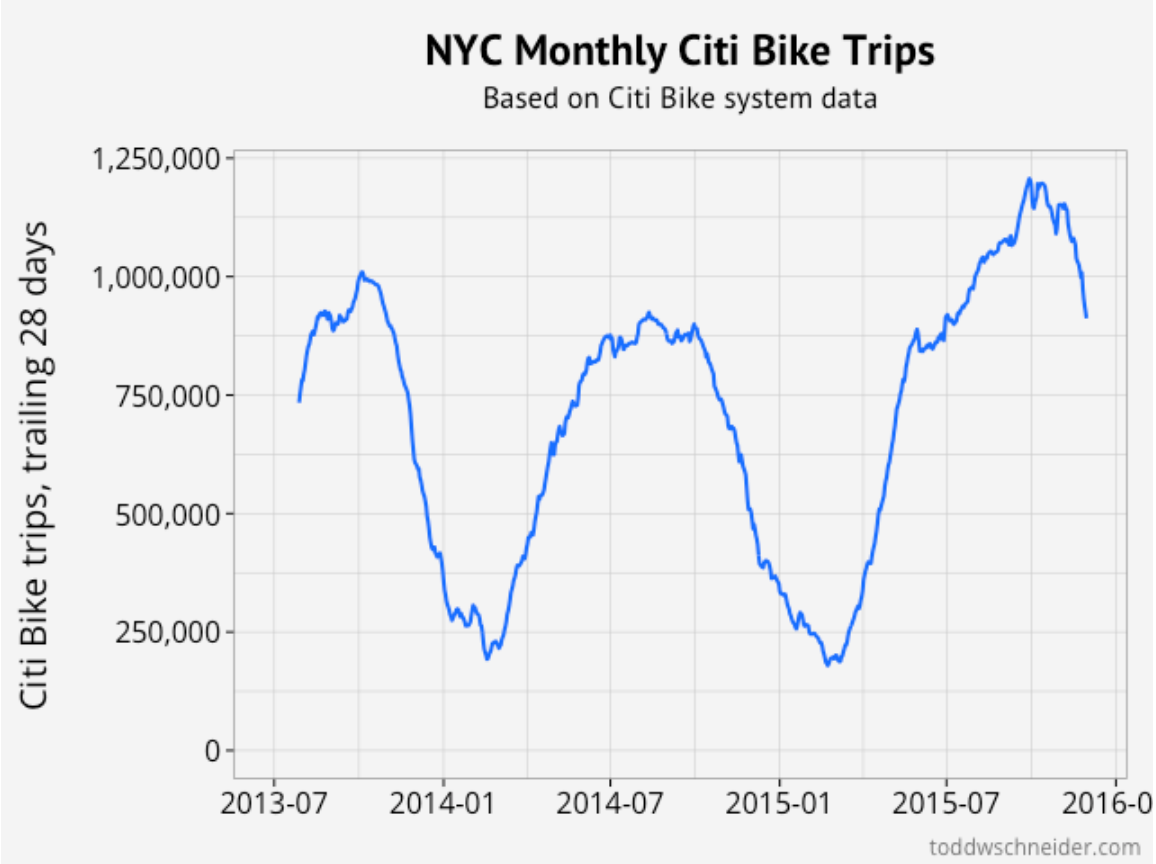
二、数据

不像出租车，公共自行车不能在城市的任意位置上、下客。相反地，骑行者们可以在城市内部有限数量的站点获得或停放自行车(https://member.citibikenyc.com/map/)。公共自行车还没有像出租车一样到处存在——2015 年大概有 1.75 亿出租车出行量，0.35 亿 Uber 出行量和 0.1 亿公共自行车骑行量——但是自行车共享系统计划在未来继续扩容。

纽约公共自行车系统使得每一次个体骑行数据都是可用的。每一条骑行记录包括：

- 骑行开始和结束的站点位置
- 骑行开始和结束的时间戳
- 骑行者性别
- 骑行者出生年份
- 骑行者是纽约公共自行车的年度用户还是短期顾客
- 使用的自行车的唯一标识

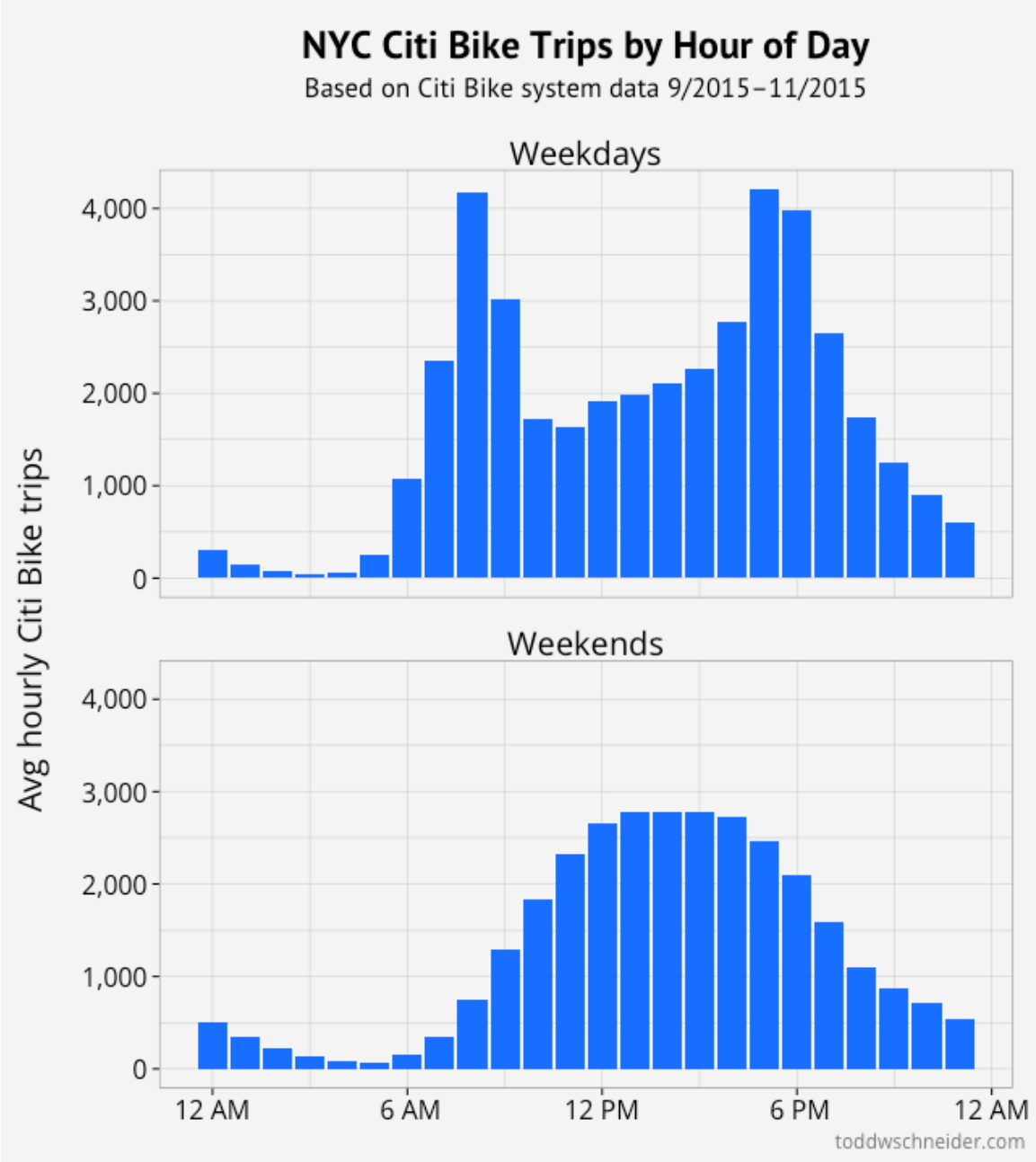
这里有一张自项目 2013 年 6 月全面启动以来每月使用量的图表：



图四、纽约公共自行车每月出行量

一点也不奇怪的是，在寒冷的冬季月份公共自行车骑行量显著减少。稍后在文章中我们将尝试量化天气对公共自行车客流量的影响。2015 年 8 月客流量的增加对应于公共自行车系统第一次大规模的扩容，这次扩容在布鲁克林区、皇后区和曼哈顿区增加了将近 2000 辆自行车和 150 个站点。

公共自行车系统在工作日的使用量要高于周末，如果我们关注一天内每小时的出行量，我们可以看到工作日骑行者主要使用公共自行车通勤去工作及回程，高峰时间在上午 8 点到 9 点和下午 5 点到 7 点。另一方面，周末骑行者的行程安排更加休闲，绝大部分周末骑行发生在下午的时间内：



图五、纽约公共自行车一天内每小时的出行量（基于纽约公共自行车系统 2015 年 9 月-11 月工作日和周末的数据）

三、年龄、性别和谷歌地图骑行时间估计的准确性

年龄和性别的人口数据可以和谷歌地图骑行导航相结合来解决一系列有趣的问题，包括：

- 公共自行车骑行者往往出行速度是多少？
- 谷歌地图骑行时间估计有多精确？

- 年龄和性别是怎么影响骑自行车的速度的？

对于每一次出行，我们根据谷歌地图上的出行距离除以出行时间代替此次出行的平均速度。这很有可能低估了骑行者实际的平均骑行速度，因为出行包括了从起点站取自行车、调整、检查电话导航或者处理其他干扰事情，以及在终点站还车的时间。

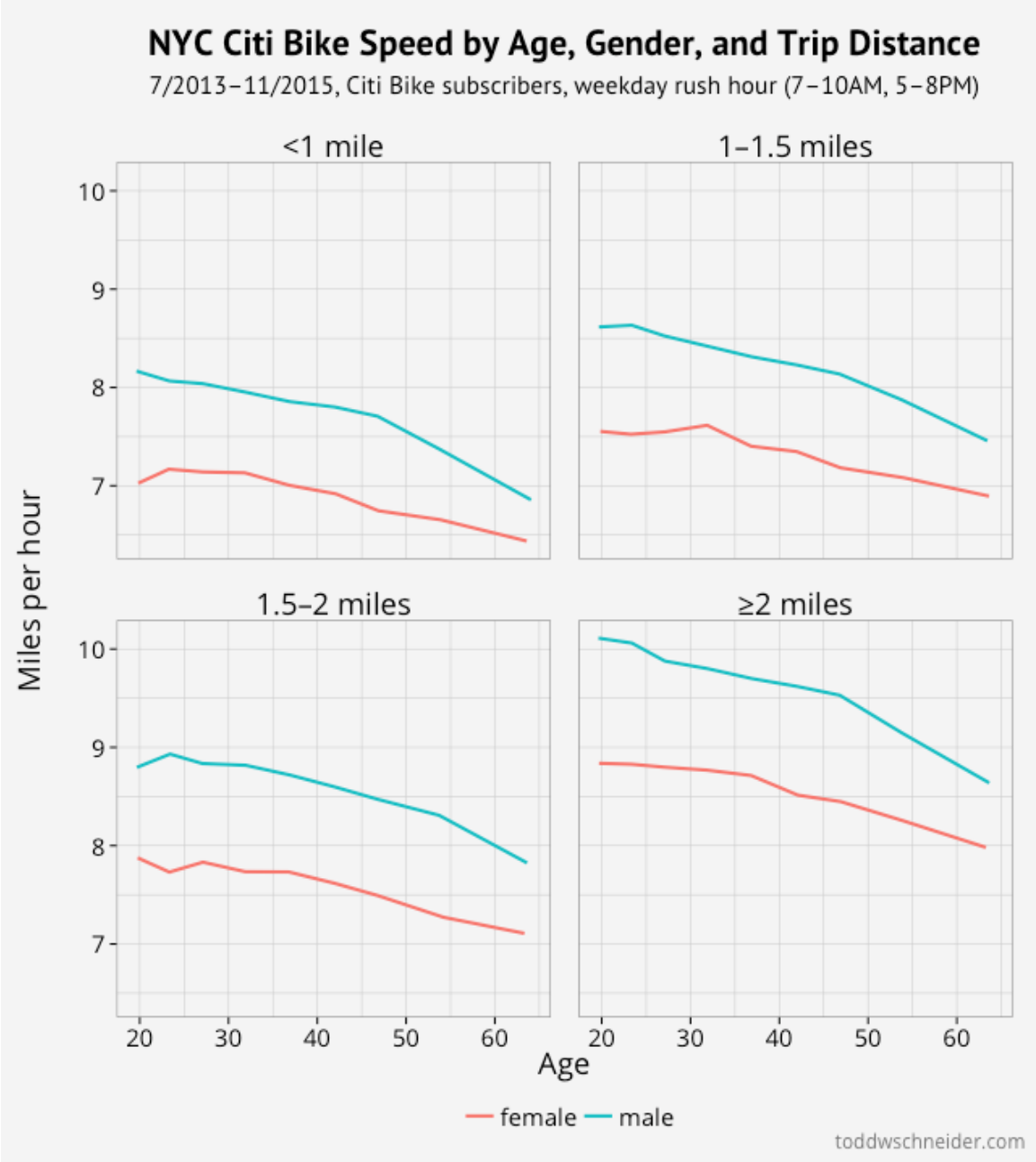
此外，假设骑行者遵循谷歌地图的导航。如果骑行者实际选择一条比谷歌推荐的路线更长的路线，那么将旅行更长的距离，我们也将低估平均出行速度。另一方面，如果骑行者选择一条比谷歌推荐的路线更直达的路线，我们可能会高估出行速度。

我们不知道任何一个个体骑行者的打算：一些骑行者可能尝试尽可能快和安全的从点 A 到点 B，而其他骑行者可能想要选择一条恰巧开始于点 A 终止于点 B 的风景优美的路线。后者几乎可以肯定不会遵循一条直达的路线，所以我们最终将为这些出行计算出一个非常慢的平均速度，即使骑行者们在整个时间内很努力的踩脚踏板。

因此，对于自行车速度的分析，我仅限于下列的出行子集，这些子集我至少可以弱弱地声称，它们更可能包括了那些试图尽快从点 A 到点 B 的骑行者：

- 工作日，排除假日
- 高峰小时（早上 7 点到 10 点，下午 5 点到 8 点）
- 年度用户
- 平均出行速度在 4-35 英里/小时（为了避免错误数据）

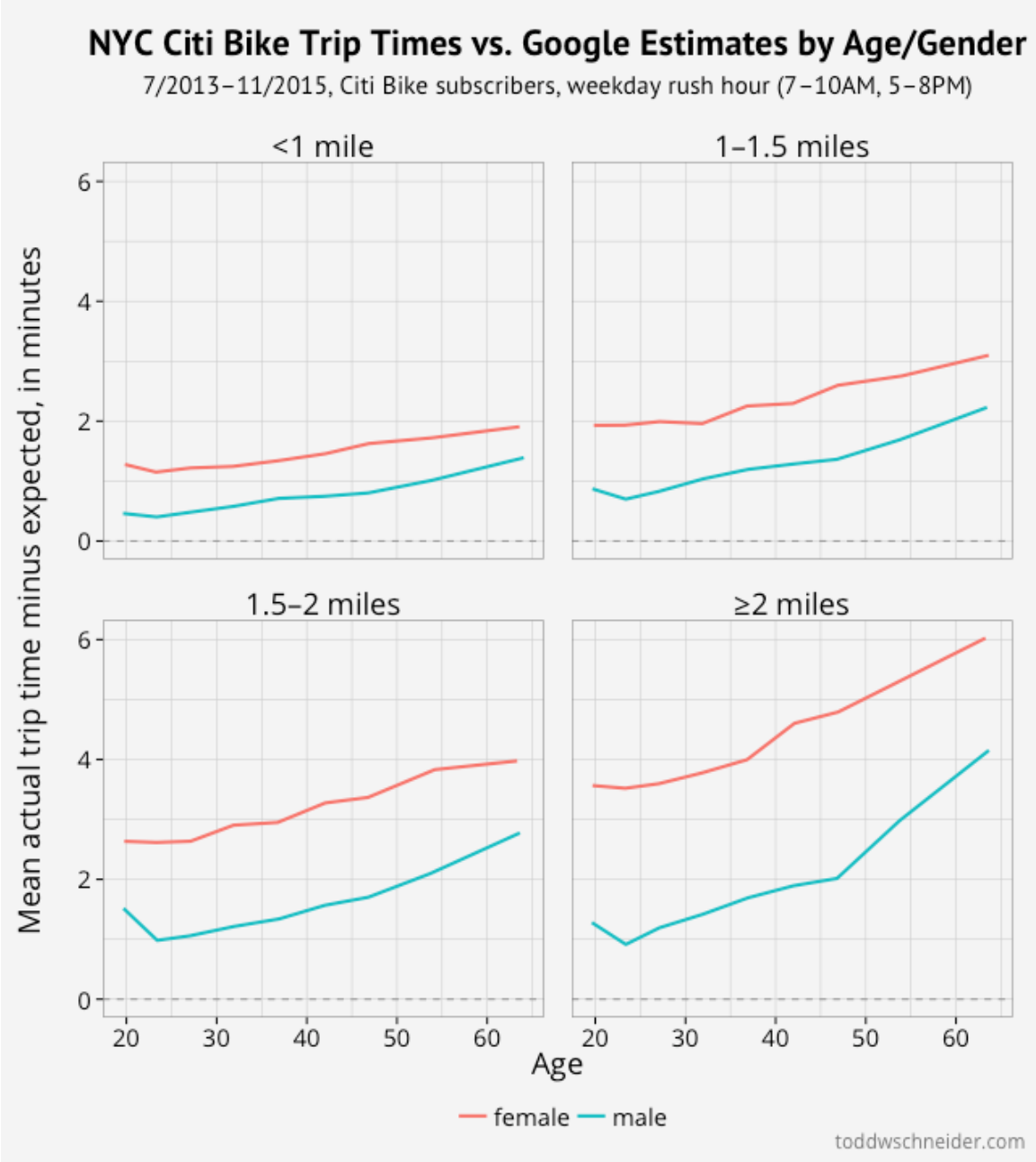
然后我根据年龄、性别、出行距离和计算的平均出行速度来定义分组：



图六、不同年龄、性别和出行距离骑行者的公共自行车速度（2013 年 7 月-2015 年 11 月，公共自行车用户，工作日高峰小时（上午 7 点-10 点，下午 5 点到 8 点））

所有这样的出行的平均速度是 8.3 英里/小时，图表表明年轻的骑行者倾向于比年长骑行者骑得更快，男性倾向于比女性骑得更快，更远距离的出行比相对短的出行的平均速度更高。

比较实际出行时间和谷歌地图估计的时间也是很有趣的。谷歌地图知道，比如，沿着一条宽阔的专用自行车道骑行的平均速度将快于沿着一条没有专用自行车道的狭窄的十字路。我利用相同的分组，计算实际出行时间和谷歌地图估计出行时间之间的平均差。



图七、不同年龄和性别分组的纽约公共自行车出行时间 VS 谷歌地图估计时间（2013 年 7 月-2015 年 11 月，公共自行车用户，工作日高峰小时（上午 7 点-10 点，下午 5 点到 8 点））

如果每个人正好花费了谷歌地图骑行导航估计的时间，我们将看到在 0 处的一系列直线。然而，每一个分组都有一个正的差值，表示实际出行时间要比谷歌地图预测的时间要慢，平均慢 92 秒。正如前面提到的，部分原因是因为谷歌地图的估计不包括在公共自行车站点花费的时间，同时我们不能保证我们数据集中的每一个骑行者都试图尽快地从点 A 到点 B。

我在 R 软件中进行了线性回归，将实际与估计出行时间之间的差值定义为性别、年龄和出行距离的函数。回归的目的并不是进行任何精确的预测——它对于更长距离出行（的时间差）的外推尤其糟糕——但更多的是为了理解每一个变量影响的相对大小：

```
1lm(formula = difference_in_seconds ~ gender + age + distance_in_miles,
2    data = rush_hour_data)
3
4Coefficients:
5Estimate Std. Error t value Pr(>|t|)
6(Intercept)             34.03         0.338846     100.4    <2e-16
7genderMale              -87.13        0.192166    -453.4    <2e-16
8age                       2.25        0.007327     306.3    <2e-16
9distance_in_miles       25.69         0.072994     352.0    <2e-16
10---
11
12Residual standard error: 213.6 on 7226125 degrees of freedom
13Multiple R-squared:   0.05475, Adjusted R-squared:   0.05475
14F-statistic: 1.395e+05 on 3 and 7226125 DF,   p-value: < 2.2e-16
```

图八、源程序 1

回归的低 R 方 0.055 再次说明数据含有大量的方差，对于任意给定的出行，这一模型不太可能产生特别准确的估计。但是这个模型至少给我们一个简单的公式，可以根据谷歌地图的估计值粗糙的估计公共自行车用户高峰小时出行需要多长时间：

Start with 34

If male, subtract 87

Add (2.2 * age in years)

Add (25.7 * trip distance in miles)

结果是实际的和谷歌地图估计的出行时间之间相差的平均秒数，正值表示比估计的出行更慢，负值表示比估计的出行更快。是的，这意味着每一年你变老，在你日常的公共自行车通勤路线上你可能会慢 2.2 秒！

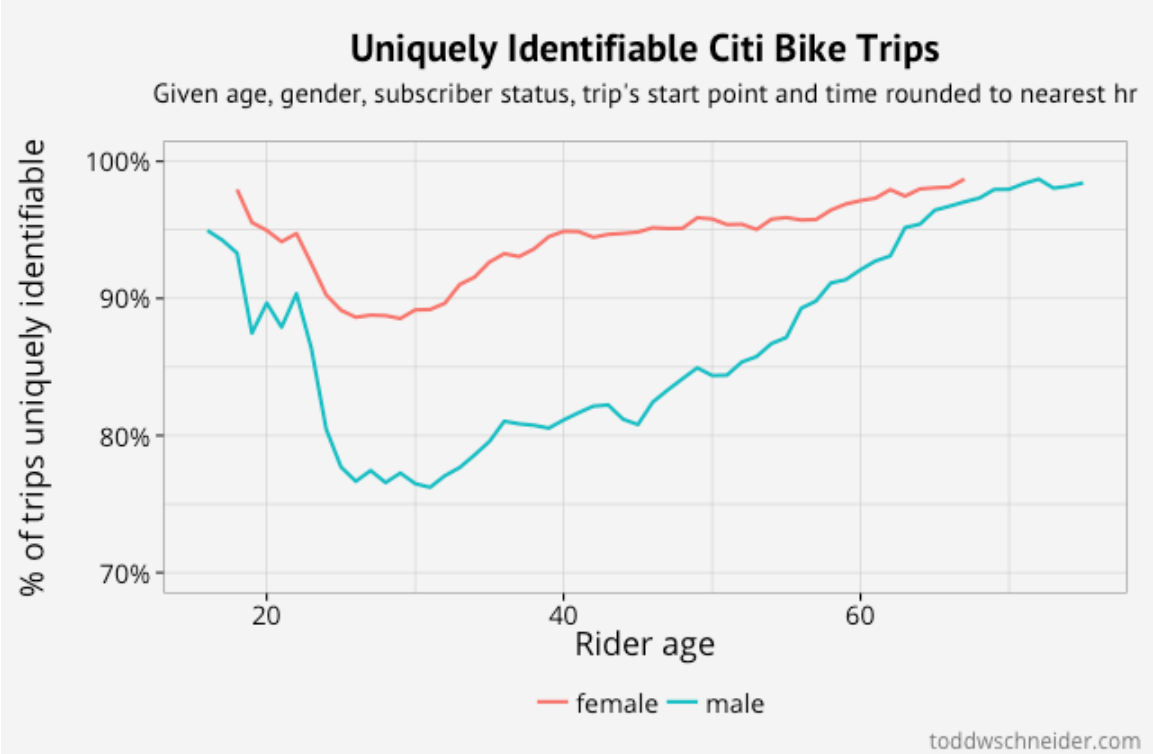
四、数据匿名是很困难的

在我关于出租车数据的文章中，我包括了一节关于数据隐私的内容（请参考其姊妹篇《交通大数据丨纽约公开 11 亿条出租车和 Uber 原始数据，大家一块来分析吧》第 10 节），并指出精确的上、下客坐标可能揭示关于人们居住、工作和社交的潜在的敏感信息。公共自行车数据没有与精确坐标相同的问题，因为所有公共自行车出行必须起止于 493 个固定站点中的一个。

但是不像出租车数据，公共自行车包括骑行者的人口统计信息，也就是性别、出生年份和用户身份。乍看上去这可能不是太暴露，但事实证明这已经足够唯一地识别很多公共自行车出行。如果你知道关于一次个体公共自行车出行的下列信息：

- 骑行者是年度用户
- 他们的性别
- 他们的出生年份
- 他们取得一辆公共自行车的站点
- 他们取得自行车的日期和时间，四舍五入到最近的小时

那么你可以唯一地识别那一次个体出行 84%的时间！这意味着你可以找出这个骑行者什么时间在哪里停放了公共自行车，这可能是敏感的信息。因为男性占据所有用户出行的 77%，唯一标识女性的骑行是更容易的：如果我们局限于女性骑行者，那么 92%的出行可以被唯一识别。那些显著年轻或年长于平均年龄的骑行者也更容易识别：



图九、唯一识别的公共自行车出行（给定年龄、性别、用户身份、出行开始站点和四舍五入至最近小时的时间）

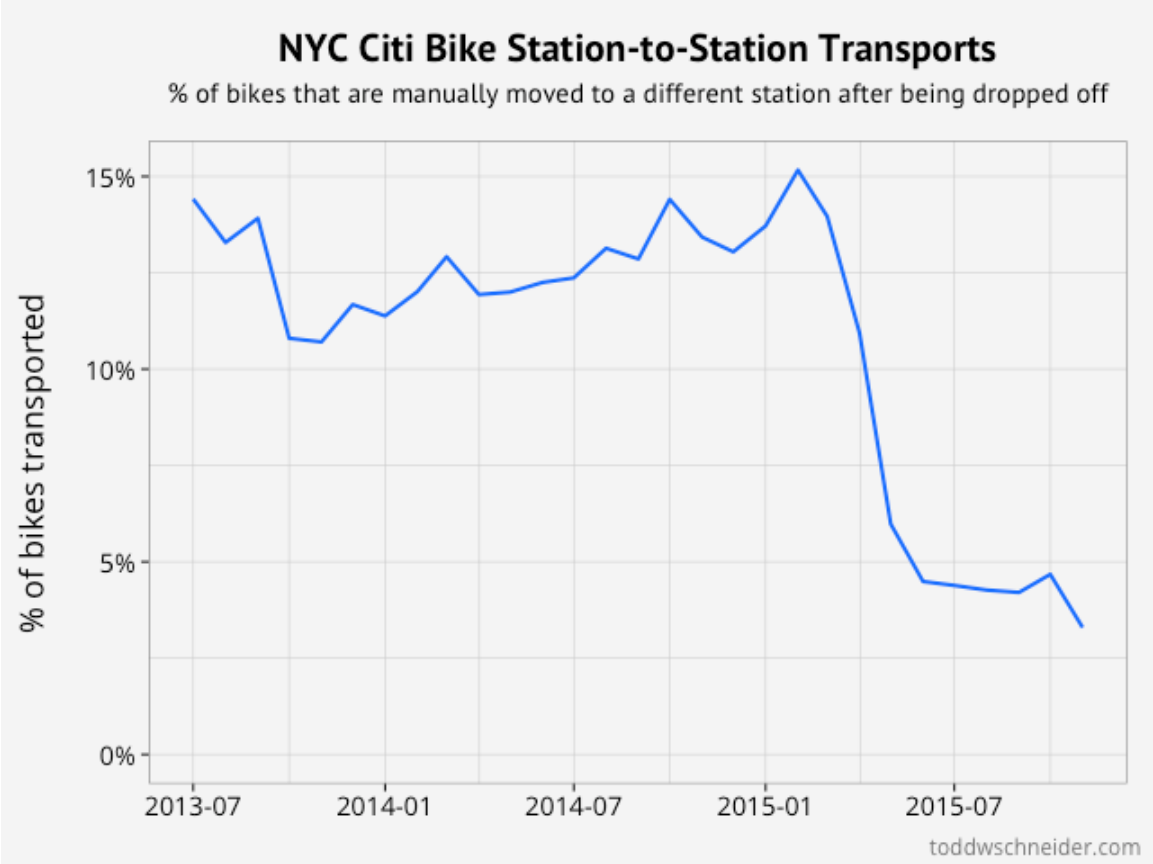
如果不知道出行开始时间到最近的一个小时，你只知道到最近的一天，那么你将能够识别所有出行的 28%，但识别的女性出行仍然在 49%。

在某种程度上这不应该太令人吃惊：Latanya Sweeney 的著名论文指出 87%的美国人口可以通过生日、性别和邮政编码唯一确定。我们可能倾向于低估从看上去有限的数据中识别个体是多么容易这样的一件事，我希望当人们决定哪些数据应该被公开时再考虑这个问题。

五、神奇的运输

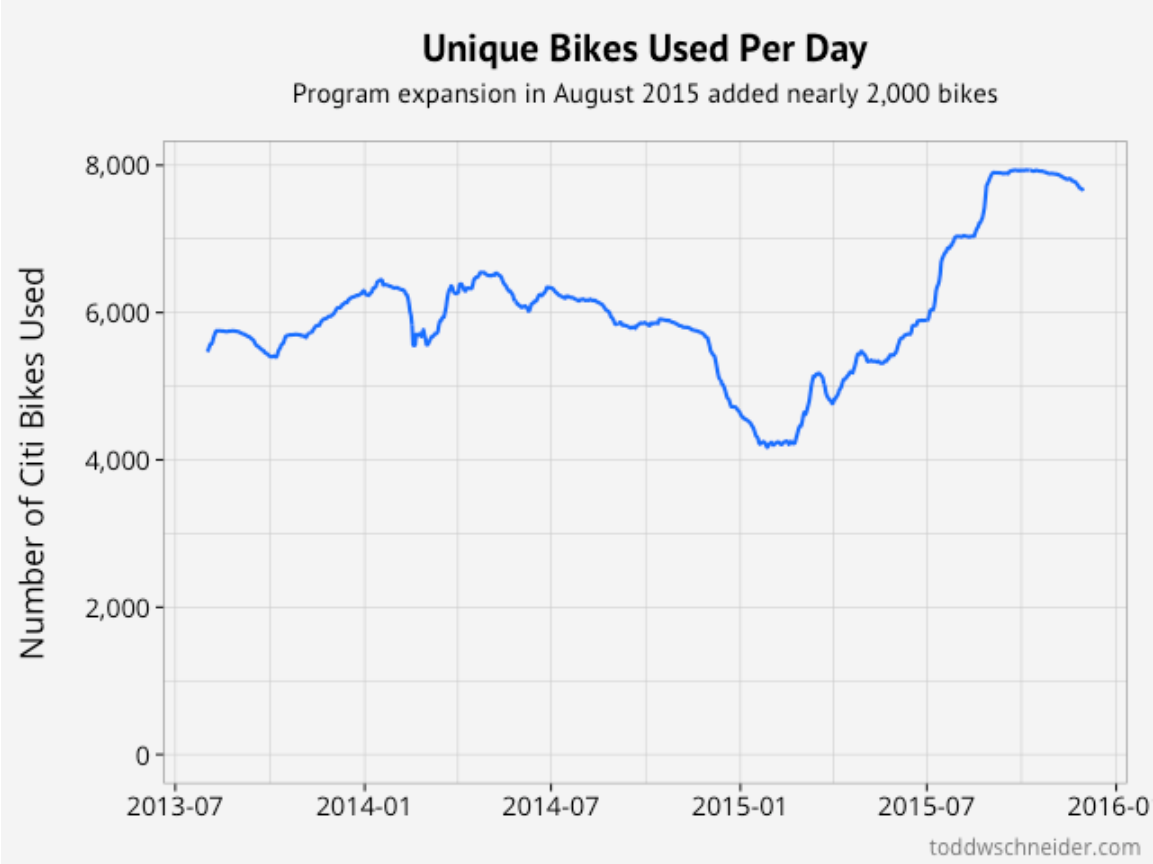
免责声明：我对运行一个自行车共享系统的物流一无所知。不过，我想象，其中一个大的问题是确保那些人们想要上车的站点有可用的自行车(即公共自行车再平衡问题)。如果站点 A 在一天开始的时候有大量的自行车，但是人们把自行车带到其他站点，并且没有人把车还到 A，那么 A 将用完所以自行车，这是糟糕的。

自行车共享运营商可以运输额外的自行车到 A 点来满足需求，但是这将花费时间/金钱，所以运营商可能想要尽可能的避免这种现象。数据让我们估计自行车“神奇地”从一个站点运输到另一个站点的频率，即使没有人骑自行车。我利用每一条自行车停放记录，计算了下一次自行车出行的起始站点与上一次出行的停放站点不同的骑行记录的百分比：



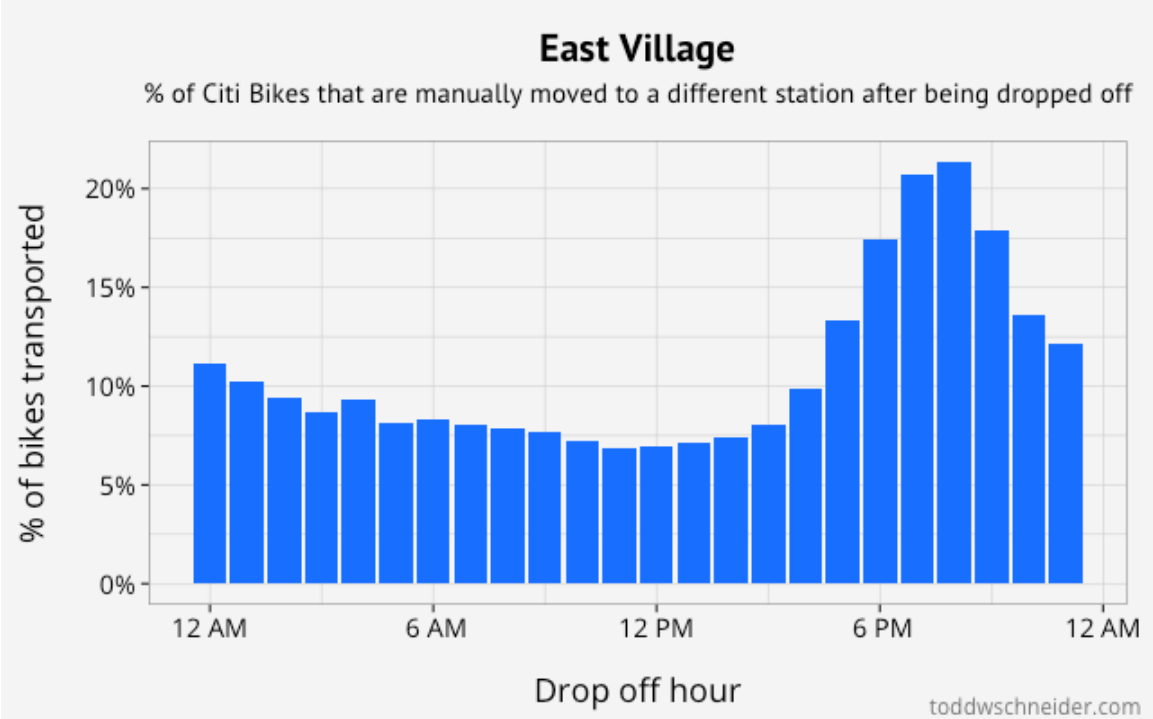
图十、纽约公共自行车站到站运输（停放后被人工运输到另一个不同站点的自行车的%）

从 2013 年 7 月到 2015 年 3 月，13%的自行车再次被使用之前不知何故被从停放站点运输到不同的站点。不过 2015 年 4 月以来，这一比例下降到 4%左右。我不知道为什么：我的第一个猜测是有更多的自行车加入到系统中，但是 2015 年 4 月正在使用中的自行车的数量没有改变。也没有自行车站点的增加或撤销，因此这似乎是一个不太可能的解释。也许是运营商开发了一个智能系统来分配自行车，这带来了更小的转移百分比？



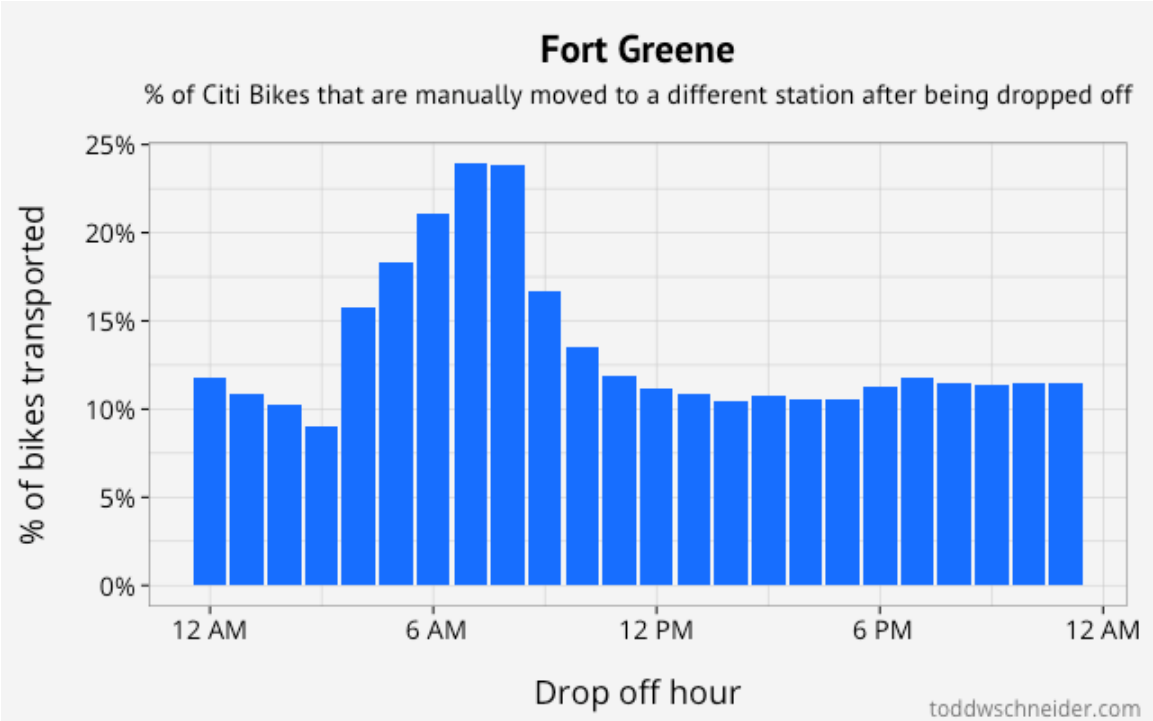
图十一、可用公共自行车数量

不同的居民区也有不同的转移模式。被停放在曼哈顿东村的自行车如果是晚上停放的，更有可能被搬运。



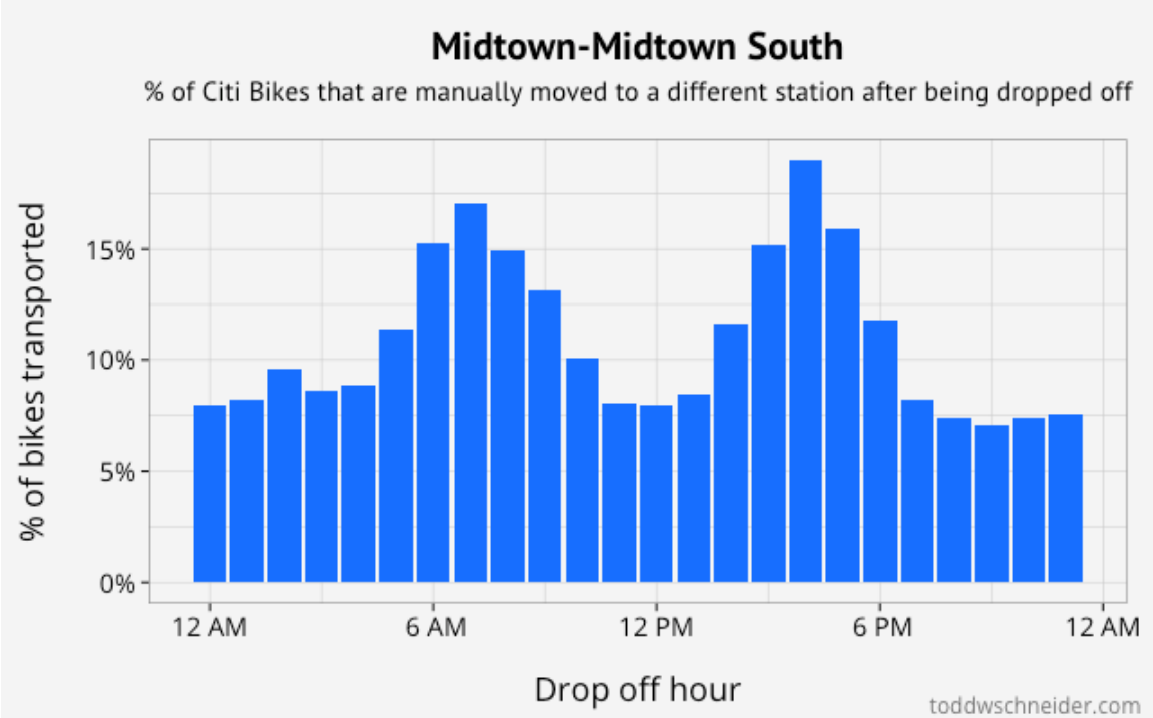
图十二、东村（停放后被人工运输到另一个不同站点的自行车的%）

然而对于早上停放在布鲁克林区格林堡的公共自行车，被搬运的可能性更大：



图十三、格林堡（停放后被人工运输到另一个不同站点的自行车的%）

在曼哈顿中城，早上或晚上高峰小时内被停放的自行车被搬运的可能性更大：



图十四、中城—中城南区（停放后被人工运输到另一个不同站点的自行车的%）

把所以的都加起来，我不确定这意味着什么，但这似乎是未来可以进一步研究的东西。公共自行车项目已经计划在 2016 年继续扩容。我想知道新的站点将如何影响转移率？

六、量化天气对纽约公共自行车活动的影响

我们前面看到夏天比冬天的公共自行车出行量更大。这并不奇怪：任何一个有一点常识的人都知道天气寒冷时骑自行车并不是很愉快的。同样地，骑自行车可能在雨天和雪天没有那么受欢迎。这引发了我的思考：根据天气预测公共自行车的客流量如何？

我从国家气候数据中心下载了中央公园的每日天气数据(下载地址 https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail)，把它加入纽约公共自行车数据，试图建立公共自行车使用与天气之间关系的模型。天气数据包括一些变量，最主要的是：

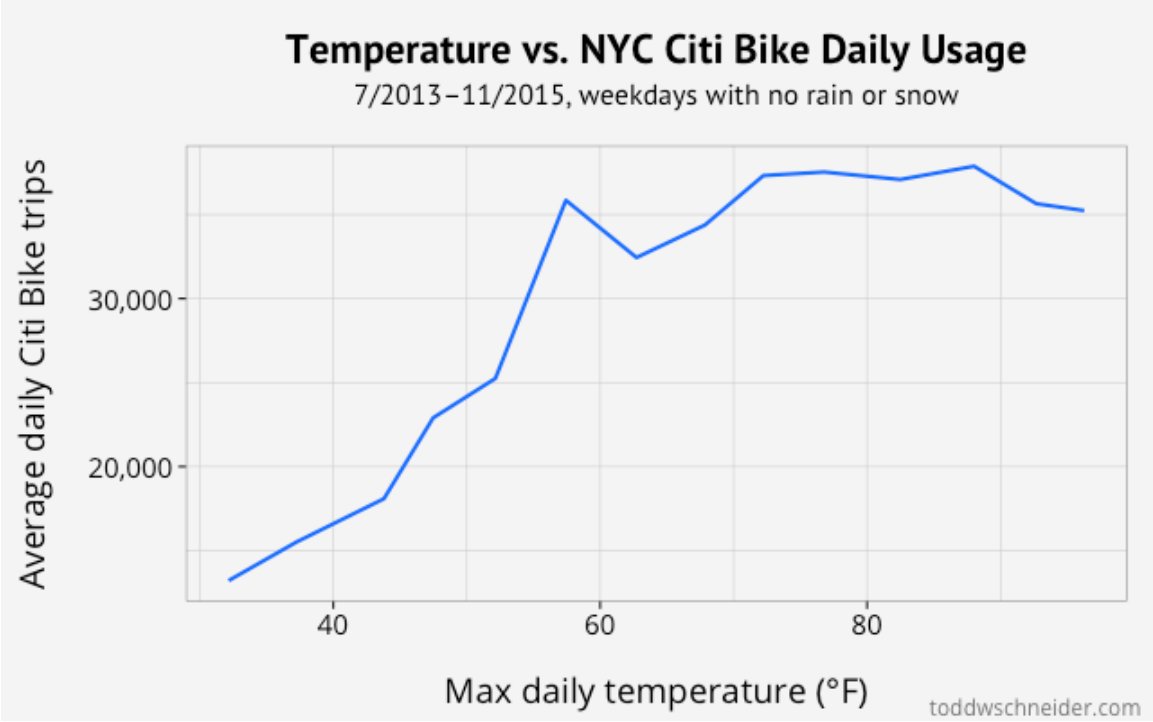
- 每日最大温度
- 每日降雨量
- 每日雪深

实际上在我开始研究这一数据之前，我怀疑线性回归将不适合这个天气模型，有两个主要的原因：

我们的因变量，日总出行量，明显是正数。一个标准的线性回归不能保证产生一个正数。

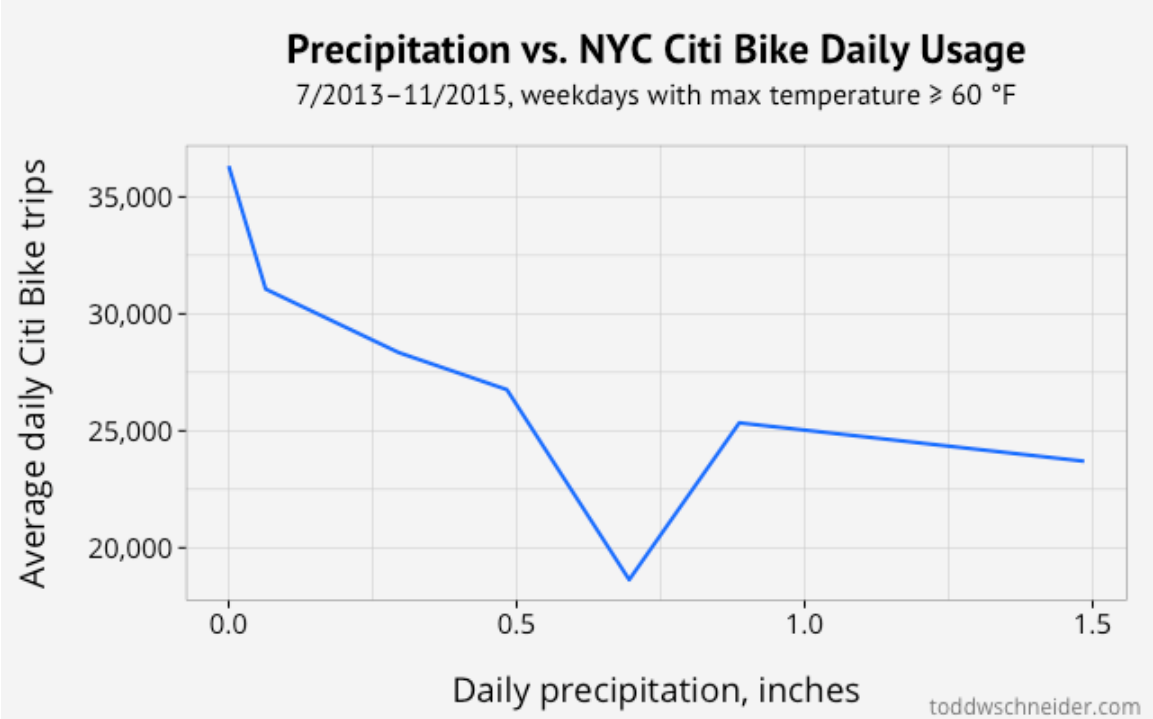
自行车客流量和天气之间的关系可能是非线性的。比如，我猜测温度在华氏 40 度到 60 度之间的日期客流量的变化可能要大于 60 度到 80 度的日期。

我们可以使用一个带有 log 对数转换的线性模型来处理问题 1，但是即使这样我们将卡在非线性问题上。让我们来确认天气和客流量之间的关系实际上是非线性的：



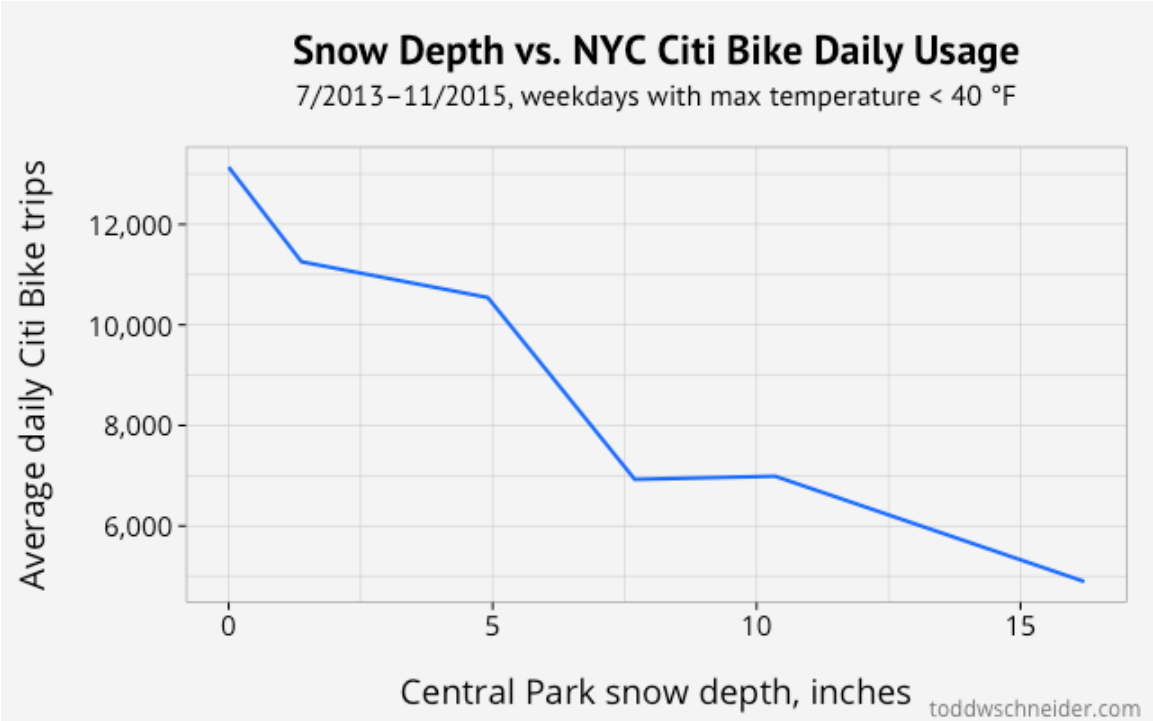
图十五、温度 VS 纽约公共自行车日使用量（2013 年 7 月-2015 年 11 月，没有雨或雪的工作日）

这张图表很清楚地表明，客流量和每日最高温度之间是非线性关系。华氏 30 度到 60 度之间客流量快速增加，但是 60 度以上客流量和温度之间的关系要弱的多。让我们看一看雨天：



图十六、降水 VS 纽约公共自行车日使用量（2013 年 7 月-2015 年 11 月，最高温度大于等于华氏 60 度的工作日）

下雪天：



图十七、降雪深度 VS 纽约公共自行车日使用量（2013 年 7 月-2015 年 11 月，最高温度小于华氏 40 度的工作日）

雨水和降雪毫无意外的与较低的客流量有关。不太清楚是不是线性关系——相比于“正常的”日期，数据集中的观测值也更少——但是凭直觉我不得不相信两者都有边际递减的影响，也就是说，没有雨水和 0.1 英寸降水的差异要比 0.5 英寸和 0.6 英寸降水的差异更加显著。

为了校准模型，不使用 R 软件的 lm()函数，我们将使用 minpack.lm 包的 nlsLM()函数，这一函数实现 Levenberg–Marquardt 算法来为非线性模型最小化平方误差。(https://en.wikipedia.org/wiki/Levenberg%E2%80%93Marquardt_algorithm)对于非线性回归，我们首先需要指定模型的形式，我选择像这样的：

$$d_{trips} = Baseline(d) + Weather(d) \quad (1)$$

$$Baseline(d) = e^{\beta_{const} + \beta_{wday} \cdot d_{weekday} + \beta_{expansion} \cdot d_{expansion}} \quad (2)$$

$$Weather(d) = \beta_{weather} \cdot \frac{1}{1 + e^{\frac{-(WeatherFactor(d) - \beta_{center})}{\beta_{width}}}} \quad (3)$$

$$WeatherFactor(d) = d_{maxtemp} + \beta_{precip} \cdot d_{precip} + \beta_{snowdepth} \cdot d_{snowdepth} \quad (4)$$

图十八、程序 2

变量 d 对于一个给定的日期 d 是已知的值，变量 β 是校准参数，大写的函数严格来说是多余的中间产物，也就是，我们可以把整个模型写在一行上，但是我发现中间函数使得公式更容易推出。让我们单步调试模型的设定，一次一行：

1. d_{trips} 是日期 d 的公共自行车出行量，即模型的因变量。我们把出行量分成两个组成部分：一个基线部分，是关于日期的函数，一个天气部分，是关于那天天气的函数。

2. Baseline(d)函数是指数函数，保证会产生一个正值。函数有 3 个校准参数：一个常数，一个非假期工作日的调整参数，一个“后扩容时代”日期的修正系数，“后扩容时代”定义为 2015 年 8 月 25 日纽约公共自行车系统增加了将近 150 个站点之后的日期。

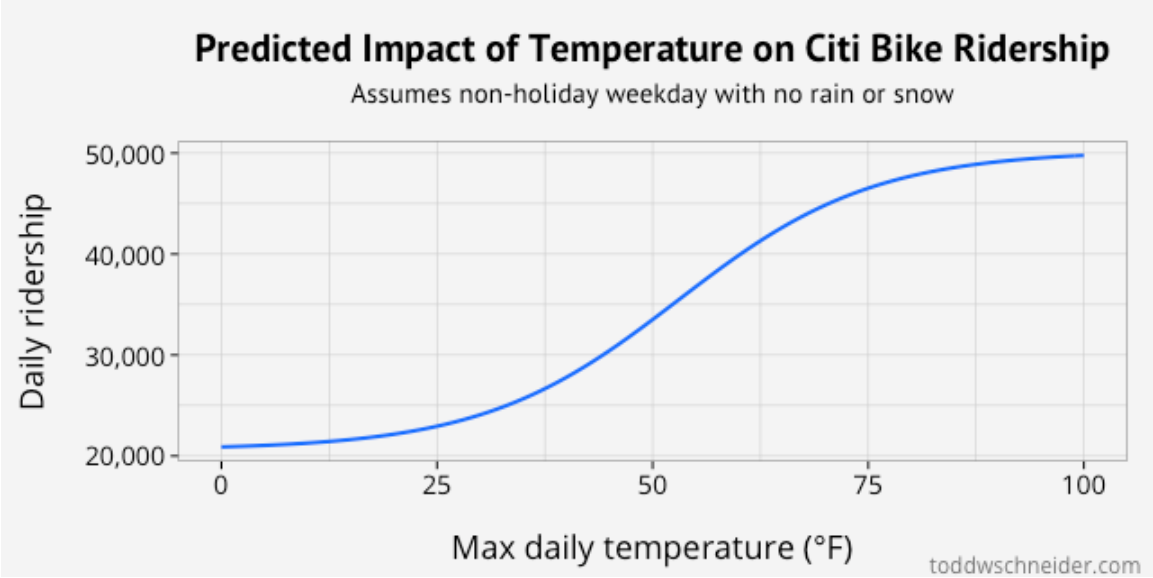
3. Weather(d)函数使用了抵押贷款的提前还款建模者最喜欢的公式：S 曲线。我欣然承认我没有“很深的”选择这一函数形式的理由，但是 S 曲线(https://en.wikipedia.org/wiki/Logistic_function)通常在非线性模型中表现很好，同时前面的温度图表看上去好像 S 曲线能拟合的很好。

4. S 曲线的输入，WeatherFactor(d)，是日期 d 的最高温度、降水量和降雪深度的一个线性组合。

这里可用的输入数据是 csv 格式（参考文献 4，请联系我们索取），你可以从这里看准确的 R 命令、输出和参数值（参考文献 5，请联系我们索取），但简单地说这个模型校准看似合理的参数。假设我们保持其他所有变量不变，该模型预测：

把日最高温度从 40 度增加到 60 度，客流量增加了 12100 次出行

而把日最高温度从 60 度增加到 80 度，客流量增加了 7850 次出行：

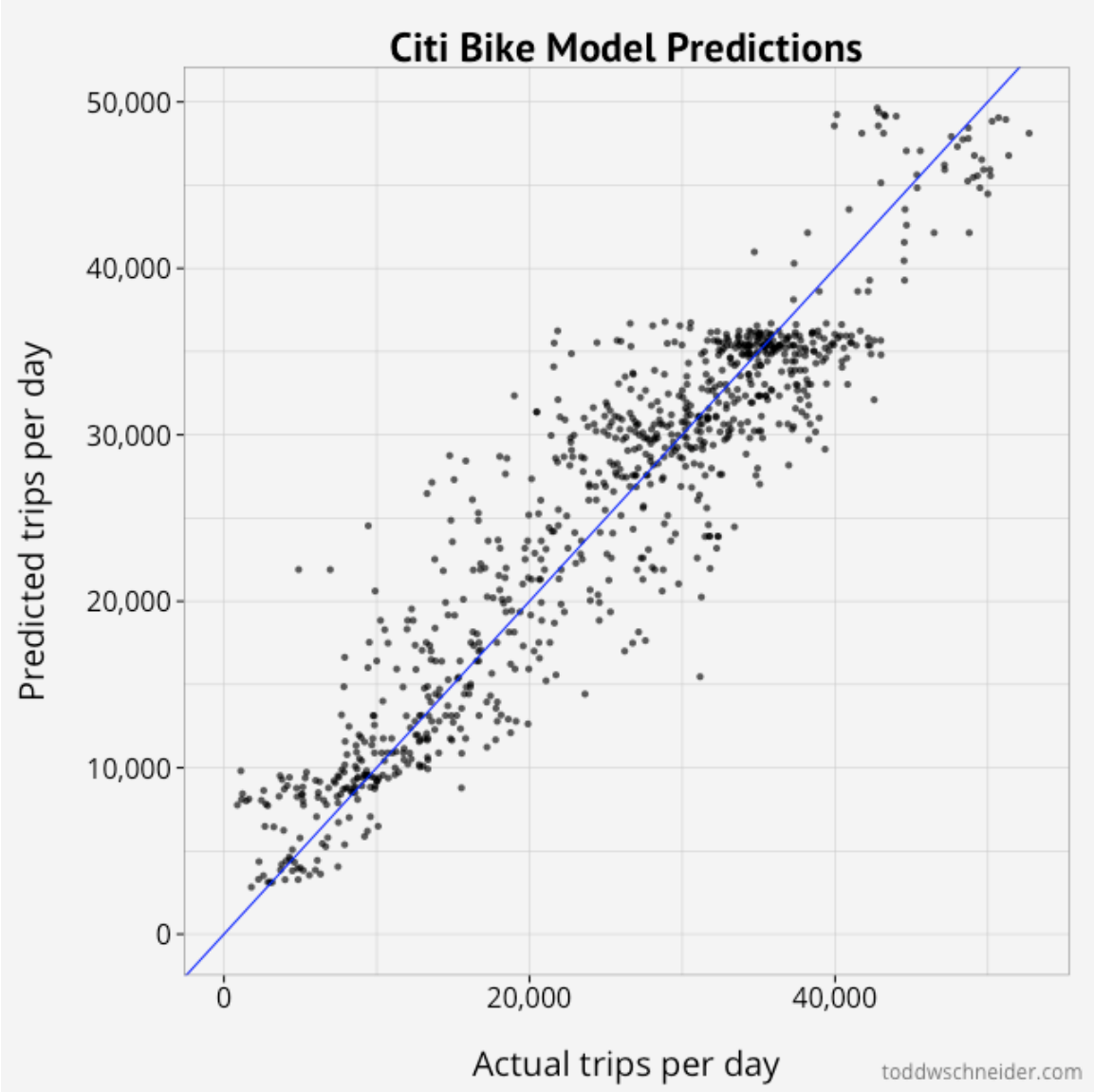


图十九、预测温度对纽约公共自行车客流量的影响（假设非假期工作日没有雨或雪）

1 英寸的降雨量和降低 24 度的温度有相同的影响

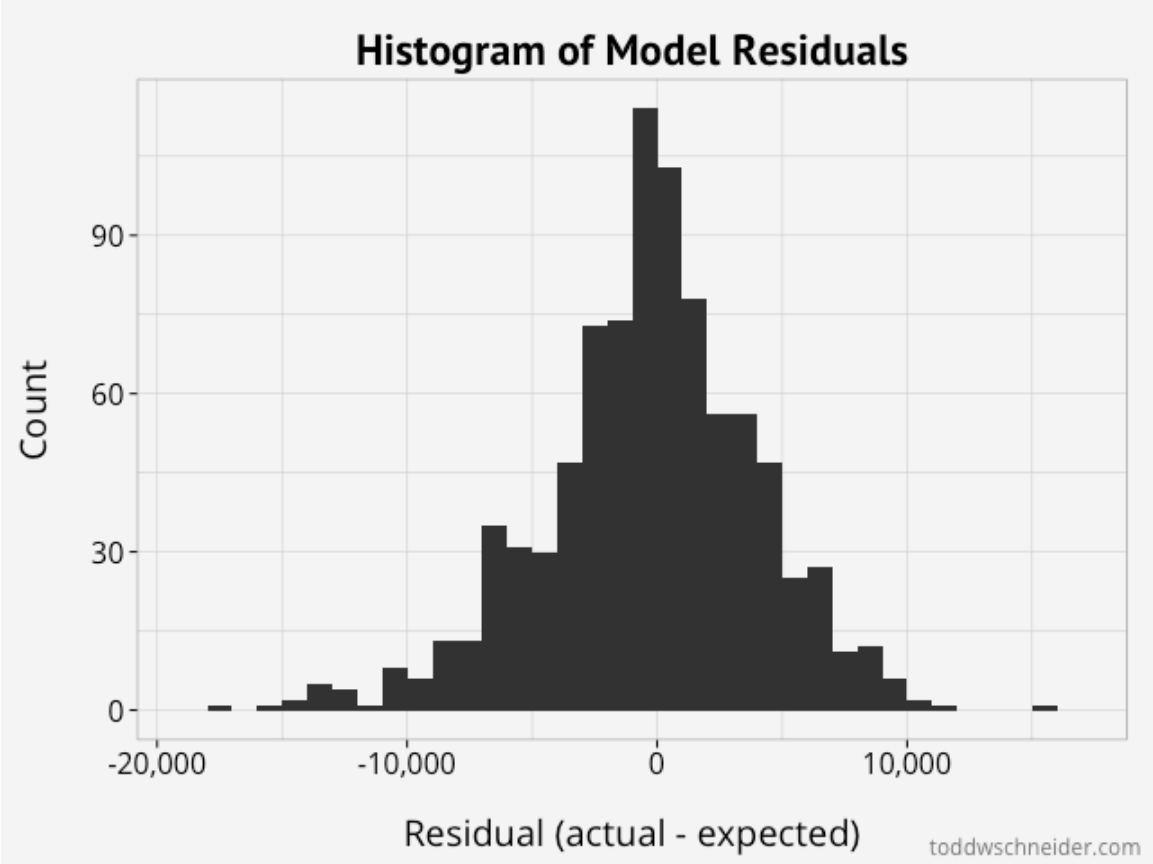
1 英寸的地面积雪和降低 1.4 度的温度有相同的影响

为了估计模型的拟合优度，我们将看一看其他的一些图表，以实际 VS 预测值的散点图开始。每一个点代表数据集中的一天，x 轴是那一天实际的出行量，y 轴是模型预测的出行量：



图二十、公共自行车模型预测

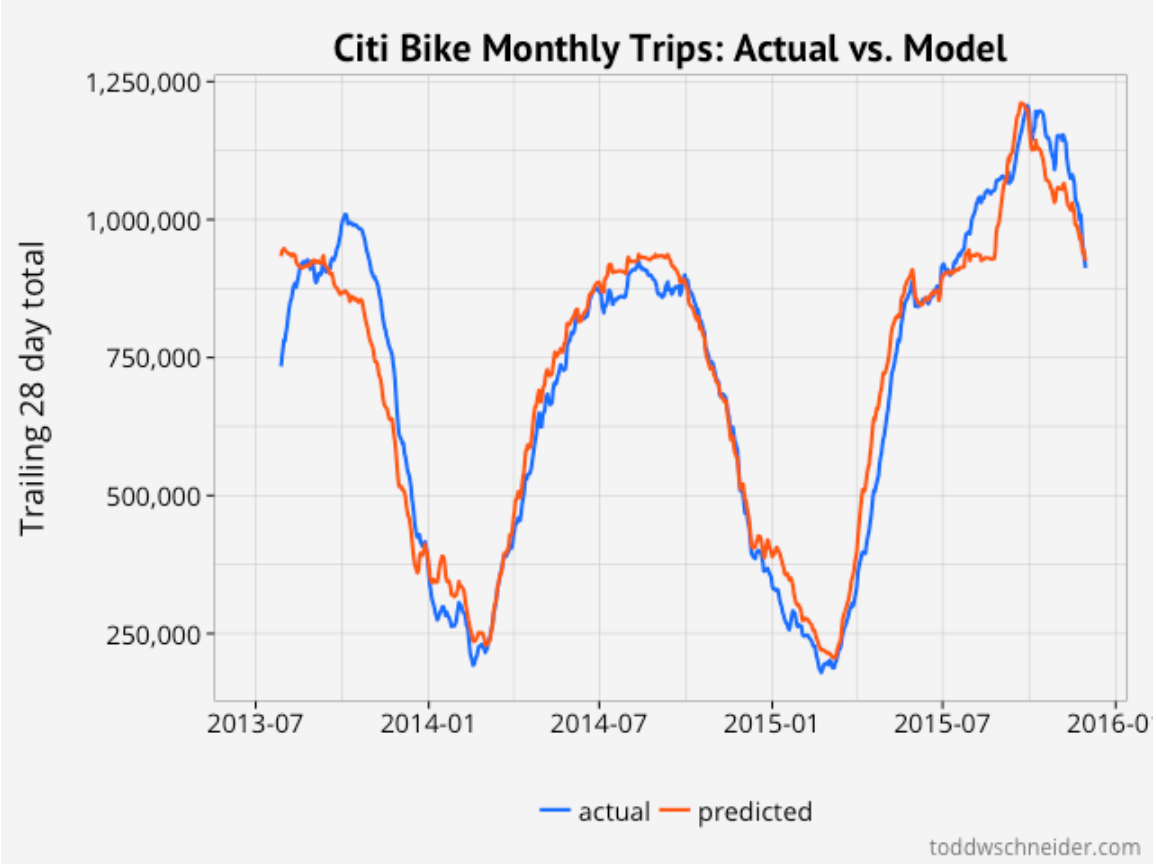
模型的均方根误差是 4.138，残差大致服从正态分布。但是残差似乎表现出一些异方差性，因为出行量较少的日期的残差有较低的方差。



图二十一、残差大致服从正态分布

“后扩容”修正系数的影响在散点图的右上角是明显的，好像对于 2015 年 8 月 26 日之前的日期在预测出行量 36000 附近有一条渐近线。理想上我们已经用公式表示了模型来避免使用一个修正系数——可能通过单个车站层面上对出行建立模型，然后集计起来——但是我们将很方便的掩饰那些。

我们也可以看一看实际 VS 预测的时间序列，集计到月总量来减少噪音：



图二十二、公共自行车月出行量：实际 VS 模型

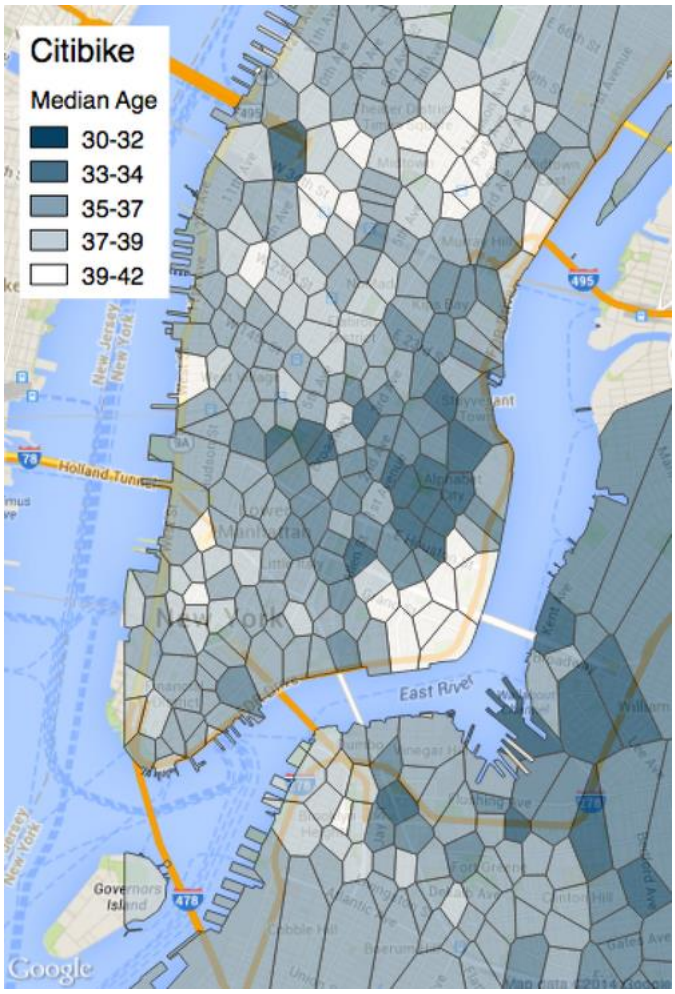
我没有声称这是一个完美的模型——它使用了有缺陷的数据，有一些不好的特性和遗漏，同时所有常见的相关性/因果关系说明适用——但似乎至少做了一件好事就是量化了温度、雨水和降雪对纽约公共自行车客流量的影响。

七、总结

一如既往地，数据集中仍然有更多的东西我们可以研究。糟糕的天气可能影响骑行速度，所以我们在测量速度和谷歌地图时间估计的时候可以把这一因素考虑进去。

I Quant NY （一个定量分析纽约公开数据的网站）的 Ben Wellington 根据站点进行了一些人口统计分析，看一看随着时间的推移是如何演变的可能也是很有趣的。

(http://iquantny.tumblr.com/post/81465368612/mapping-citi-bikes-riders-not-just-rides)



图二十三、I Quant NY 人口统计分析

我想知道关于单个站点层面客流量的建模，尤其是未来新加入的站点。增加新的站点很容易影响现存站点的客流量——甚至不清楚影响是积极的还是消极的。新的站点可能从附近的其他站点调拨出行量，可能总的客流量不会增加多少。但是也有可能新增加的站点和现存站点有协同作用：想象这样一个场景，一个进入地铁不方便的居住区有了一个公共自行车站点，那么位于最近地铁站的现存的公共自行车站点使用量将激增。

这里也可能有大量的关于比较纽约公共自行车数据与出租车和 Uber 数据的分析可以做：和出租车出行相比，哪些居住区有最高和最低的公共自行车出行比例？高峰小时交通中，是否有选择公共自行车比选择出租车要快的通勤出行？哎，这些分析可能不得不等一等…

八、GitHub 存储库

在 nyc-citibike-data repository 中有下载、处理和分析数据的脚本。csv 格式的天气分析的原始数据（日出行总量加天气数据）包括在内，以防你不想下载全部的数据。