Mobike 单车数据可视化项目 3.0

总结：

1. 根据用户的骑行次数统计，90%的用户（一般活跃用户和较活跃用户）集中于骑行 2-14 次的区间。

2. 用户人数在 8 月内稳步增长，增长的速度从高到低分别为不活跃用户>一般活跃用户>较活跃用户>活跃用户。

3. 在 24 小时之内，四组用户均呈现了类似的用户行为，且早晚高峰的骑行明显增加。但早晚高峰的骑行行为和热门位置也略有差异：不活跃用户在早高峰的骑行趋势高于其他三组用户。四组用户在早高峰的骑行偏向高频多次，且延地铁站站点骑行的行为较为明显。晚高峰时在中心城区的热门骑行区域更广，延地铁站方向的骑行没有早高峰那么明显，骑行时长也高于早高峰。

4. 上海的热门骑行点基本围绕浦西中心城区内的最密集的地铁站站点附近。而晚高峰的最热门骑车区域在五角场附近。

故事核心探索与设计：

根据上次 reviewer 关于不能只是进行数据探索，而是需要深入探究和深层思考某个实用核心问题的建议，同时在我已有可视化的基础上，我决定进一步深挖用户习惯的这个问题核心。

从而围绕用户习惯这个核心故事，按思路顺序做出了以下 7 张图。

故事的思路和可视化的核心思考如下：

- 第一张图，为建立在上个版本的已有的骑行用户分布次数统计，根据评审的反馈，重新划分了 bin size，修改了正确的聚合参数。为了探索用户习惯，我根据用户使用的分布情况，对用户分了四组。我的预期是这四组用户的习惯应该不同，那么后期可以针对不同的习惯，针对性的提升业绩。

骑行次数	用户分类	人次	比例
0-2	不活跃用户	1194	7%
2-8	一般活跃用户	10447	62%
8-14	较活跃用户	4734	28%
14-26	活跃用户	512	3%

- 第二张图， 针对这四组用户分别绘制当月内每天的用户数，用户骑行总时长，用户平均骑行时长。

  四张图中都显示了用户人数在 8 月内稳步增长，增长的速度从高到低分别为不活跃用户>一般活跃用户>较活跃用户>活跃用户。其中，不活跃用户的人数从月初的十几位用户，非常缓慢的增长，然后突然在 8 月下旬迅猛增长，直接在 8 月 31 日当天增加到了 205 人。同样的现象也出来在了一般活跃用户中，8 月整个月从最初的 538 人到 31 日的 2714，用户数翻了 5 倍。较活跃用户则翻了一倍，而活跃用户数接近翻倍。上次报告已经提到，摩拜单车在上海于 2016 年 4 月下旬开始投放，因此到 3 个月后的 8 月，摩拜单车仍然是一个新鲜事物，从无到有的过程中，用户的稳步增长是可以想见的。而且，此时已经作为摩拜单车的活跃用户，在单车使用上已经趋于稳定和成熟。反而是不活跃或一般活

跃用户，在用户注册激活，增加单车使用率等方面，都可以考虑进一步用运营手段来加强或提升。

值得注意的是，在不活跃用户组中，8 月 6 这天，只有地量的 8 名用户，用户平均骑行时长却有 60 分钟/人的异常值。较活跃用户组中在 8 月 5 日这天，也有一个平均骑行时长 25 分钟的值，异于平日。我猜测是由于异常锁车造成的。在上次的报告中已经提到了发现有很多异常锁车的异常值。于是我仅仅筛选出了骑行时长在 50 分钟以内的订单，重新作图。修正后的图为故事中的图 3。

- 第三张图，经过修正之后，不活跃用户组的平均骑行时长从之前的 20-60 分钟的区间，降到了 10-20 分钟的区间内。而另外三组用户的平均骑行时长也都在 10-17 分钟之间稳定波动。这一方面是跟数据样本较多有关，另一方面也是由于较活跃用户的单车的使用习惯趋于成熟和稳定。

- 第四张图，在上一版可视化的基础上，分用户组，显示了 24 小时之类每个时段的骑行单车数，骑行总时长，和用户平均骑行时长。

  针对四组用户，均明显的表现出 0-5 点之间，骑行次数很少，6 点骑行开始增加，早晚高峰的非常明显。早高峰集中在 7-8 点，晚高峰集中于 15-17 点之间。中间两组的骑行单车数趋势非常类似，均呈现早高峰比晚高峰的骑行单车数和总时长都略低。活跃用户的早晚高峰的骑行单车数趋于一样。不活跃用户在早高峰的骑行单车数反而高于晚高峰，虽然骑行总时长低于晚高峰。这可能是由于不活跃用户（每月 0-2 次）会考虑把每月骑车的配额用在早间更赶时间的时候。

  四组都呈现出早高峰的用车辆不小，但是总骑行时间和平均骑行时间都低于晚高峰时段。原因可能如下：早高峰时段大家都赶时间，骑行速度可能快于晚间，而且大家会更倾向于坐地铁，则会骑短途到附近的地铁站。同时坐地铁的前后都需要骑行，就造成了多次短时间的骑行。因此我预期上海地铁站站点位置应该是热门的单车起始点，这个我会在第六张热力图上探究。而晚间大家对时间没有那么敏感，速度相对较慢，同时跟地铁相比，也许更偏向骑行或其它交通方式。

  值得注意的是，四组用户在 0-5 时这个最不活跃的时间段，平均骑行时长居然高于早高峰。我怀疑还是由于错误锁车所致，尤其会让单车在晚间连续计时。于是我只选取了骑行时长在 50 分钟之内的订单，重新做了图 5。

- 第五张图，在修正了异常值之后，除去不活跃用户，其它三组用户的平均骑行时长都明显下降到了 10-17 分钟的区间。不活跃用户的略有下降，但是由于样本较少，波动仍然很大。

- 第六张图，尤其刚次提到了，我预期热门地铁站点应该是在早晚高峰时单车的热门起始点。于是想要叠加上海地铁站点的信息到地图上。我从以下网站获取了上海地铁站的坐标，https://blog.csdn.net/a364572/article/details/50483568 ，并在 jupyter notebook 中进行了数据处理，存入 shanghai_metro.csv 的文件中（见附件），用于导入 Tableau 进行可视化。

  遗憾的是，我查看了很多资料，也尝试了很多方法，仍然无法把这两份来自不同数据表的坐标叠加在一张地图中了，于是就只能采用画中画的方式。而且由于 Tableau 不能呈现上海的区域地图，我只能同时配合百度地图查看具体的上海区域。

如果选取早晨 8 点这个时段，红色热力区域集中于地铁站站点最集中的中心城区，即内环的徐汇区一路向北到五角场附近。同时在中心城区外，呈现红色热力点也呈现带状发散。这些条带明显是地铁站的走向。说明确实早高峰在中心地铁站，还有城区外部的地铁站点，是高频的单车起点。用户早高峰进城前会骑车到达地铁站，进了中心城区，下了地铁也会选择单车，以节约时间。这样也就解释了为什么早高峰的单车使用数不亚于晚高峰，但是平均骑行时间要低于晚高峰。

再选取 18 点这个晚高峰时间，这次红色热力区都仍然主要集中于主城区，但是这次热力区域要大于早高峰的热力区域。主要热门区域为黄浦江以西，从上海南站一路向北到达五角场。五角场区域为最热门的骑行地点。带状散开的红点区域，也都符合地铁站点的分布，但是带状分布的没有早高峰那么明显。这说明晚高峰时，大家使用单车的区域虽然都集中于中心城区，但是比早高峰更散，没有那么集中地依靠地铁出行。

- 第七张图，仍然保留了上一版本的全部用户的骑车统计和地图位置查询。根据审阅者的反馈，这张图的指导意义不大。故删除。

以下链接为最新的 story 4.0 版本。同样的链接也能找到前三个版本。

https://public.tableau.com/profile/ting.wang#!/vizhome/MobikeProject/MobikeStory4_0

反馈汇总

通过资深互联网人士，成都地区前滴滴市场经理和运营经理的意见反馈，针对 story 初版里的 5 张图，分别收集意见如下。

Story 初版可视化编号	反馈意见	针对反馈对可视化所做的修改
1. 骑行密度图	每个骑行点可以考虑扩大区间，做成热力图，有颜色变化，更便于区分高密度高频区域	改成热力图，密度大订单多的地区为红色，低频的区域为蓝色
2. 每日及每时段骑行时长和单车数量图	每日骑行时长和用户数随日期的变化图，不应该使用面积图，应该如下图一样，对每日用户数做柱状图。 每日骑行图可以看出用户和骑行时长在稳步上升。应该对每天单用户平均骑行时长做可视化，进一步探究稳步增长的原因，是因为用户变多还是单用户骑行时长增加了？	每日用户数改为柱状图。同时加入了用户每日平均骑行时长。 由图课件，用户每日平均骑行时长并不随日期增加。可见骑行总时长的稳步增加，是因为用户数的增长。

3. 统计箱线图	没有意见。	
4. 活跃用户排名（只筛选了前 20 名）	这个图对运营方没有太多价值。这些活跃用户数目不多，为特殊值，反而应该剔除掉。应该做直方图，统计用户每个月骑行次数的频率，从而了解中间那些高频次的用户行为，以及频次分布是否为正态分布。	删除了活跃用户排名图，改为用户当月骑行次数的直方图。
5. 最活跃用户的骑行次数和位置查询	最活跃用户对运营方没有太多价值。可以考虑做成所有用户的每时段骑行次数和位置查询，起码对客服端有价值，可以了解每个用户的一些骑行高频时段和轨迹，从而了解他们的骑行习惯。	改为针对所有用户的每时段骑车次数，和骑行起始位置的查询。

	反馈意见	**针对反馈对可视化所做的修改**
关于第二版的故事核心的设计	在现有的初版的可视化的基础上，要把分散的图集中讲一个故事，那么最好围绕用户习惯来着手。可以对用户进行分组，看看分了组的用户表现出来的骑行情况有些什么差异。如果有差异，那么可以深入探讨差异背后的原因，要么是用户习惯，要么跟地理位置有关。	1. 将用户分成了四组 2. 分组查看用户的每天和每时段的骑行趋势和现象，并修正了异常值，让骑行现象更容易被理解 3. 针对用户骑行习惯的差异，再热力图上加入了上海地铁站点分布，用于验证我的关于早高峰在地铁站前后进行骑行的猜测。

延伸和建议

通过以上故事的讨论，有如下一些发现和建议：

1. 除了非活跃用户，其它三组用户在单车的使用行为上趋于稳定，尤其是在 24 小时内的用户行为是跟用户自身的需求相关的，基本不会随时段发生变化。所以针对这类用户在使用时段上的促销，对用户的刺激较小，没有太多价值。

2. 早高峰，用户集中的依靠地铁出行。在地铁站前后的两段行程中，都需要使用单车，造成了高频短时的用车需求旺盛的情况。但由于晚高峰时，用户出行并没有那么依赖地铁，或者用户选择更长途的骑行，所以导致很多单车并没有在晚间回到地铁站点。为了保证次日早高峰的单车需求，那么在单车并无过剩的情况下，地铁站沿线的布点和准备，可能需要人为调度车辆，这个工作量和成本是不能被忽略的。

3. 因为本数据集记录的是 2016 年 8 月的单车数据，而摩拜在 2016 年 4 月下旬才开始在上海开始投放业务，因此 8 月的数据也还是公司的业务初期的数据，大家对新事物的接受都还没有成型或者单车技术问题较多，从而造成无法锁车的很多异常值，同时我在分析时只分析了单车起始点，并没有考虑骑行轨迹和终点，因此本次分析也具有局限性。后期可以更完整收集数据来分析，从而对提升业绩提出更好的建议。

参考资料

https://stackoverflow.com/questions/17951820/convert-hhmmss-to-minutes-using-python-pandas

https://www.tableau.com/about/blog/2016/7/how-create-density-maps-using-hexbins-tableau-56511?signin=46f6c64a0c11c268f770bca804fd5b20

https://kb.tableau.com/articles/howto/histogram-of-aggregated-values

https://blog.csdn.net/a364572/article/details/50483568

https://blog.csdn.net/xxceciline/article/details/80405129