To: Paul Alley (CoS IT) & Mary Catherine Snyder (SDOT)
From: UW Team - Shreya, Sahil, Allison, Nathan
Date: April 5, 2019
Subject: Highlights from Exploratory Data Analysis

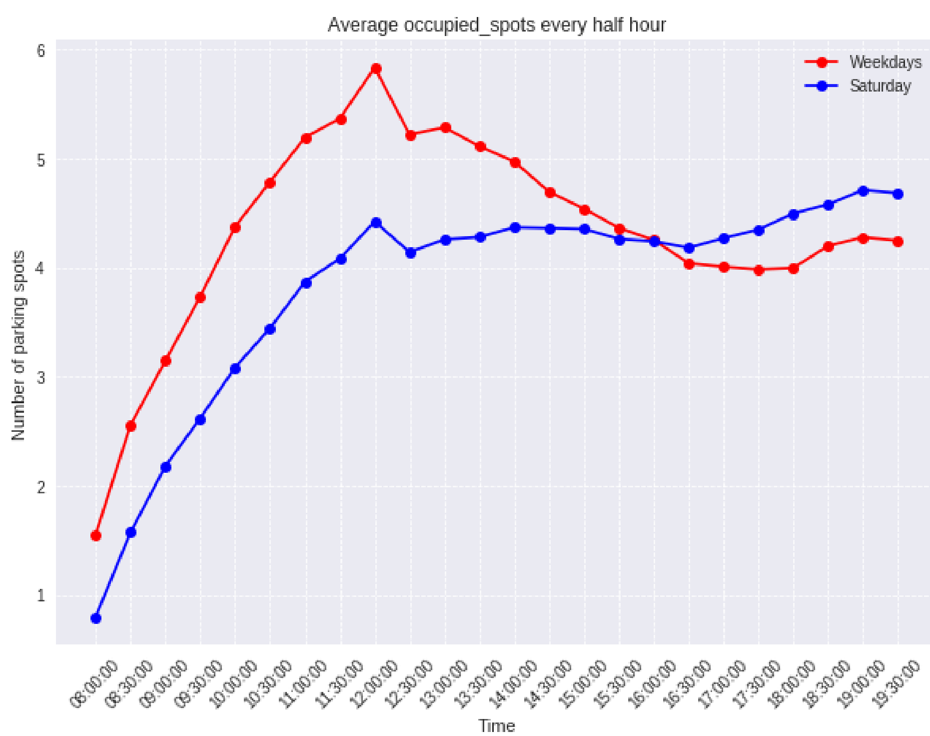
This memo summarizes key findings from the team's exploratory data analysis (EDA). Note that the domain focuses on the Belltown North neighborhood of Seattle in 2017 and 2018.

1. **The final data model contains** 51 **variables (columns)**, which represent fields collected from the data sources outlined in data pipeline documentation. For clarity and reproducibility, EDA documentation includes a list of these variables with their respective definitions.

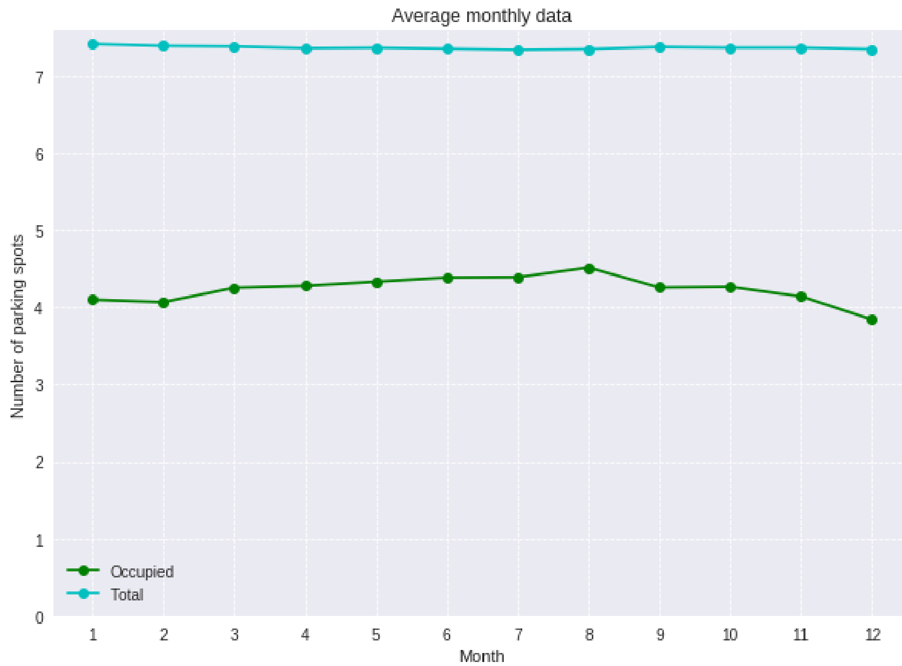
2. **The final data model (which combines all data from 2017 and 2018) contains** 2242564 **records (rows).** 2017 has slightly fewer records compared to 2018 with 1083460 and 1159104 respectively.

3. **Overall, the AUTOMATIC transaction data shows that, on average, 57.55% of spaces were occupied** out of total spaces across both years. This average occupancy increased 1.33 percentage points from 2017 to 2018. **Overall, the MANUAL survey data shows that, on average, 80.2% of spaces were occupied** out of total spaces across both years. This average occupancy increased 2.24 percentage points from 2017 to 2018.

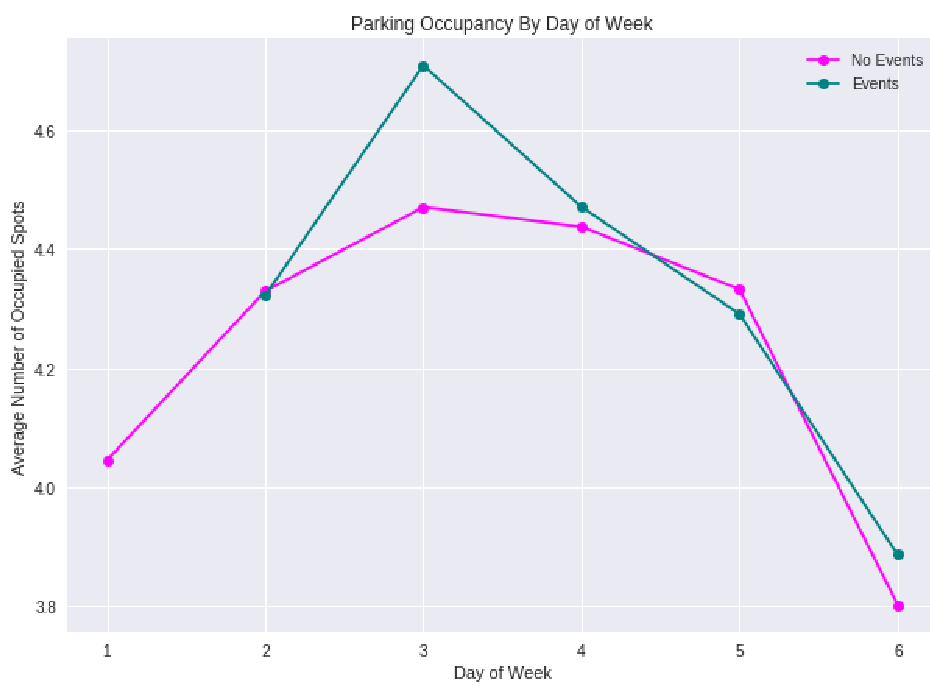
4. **Average** occupied_spots **(AUTOMATIC transactions) generally peaks weekdays around 12pm and weekends around 7pm across both years.**

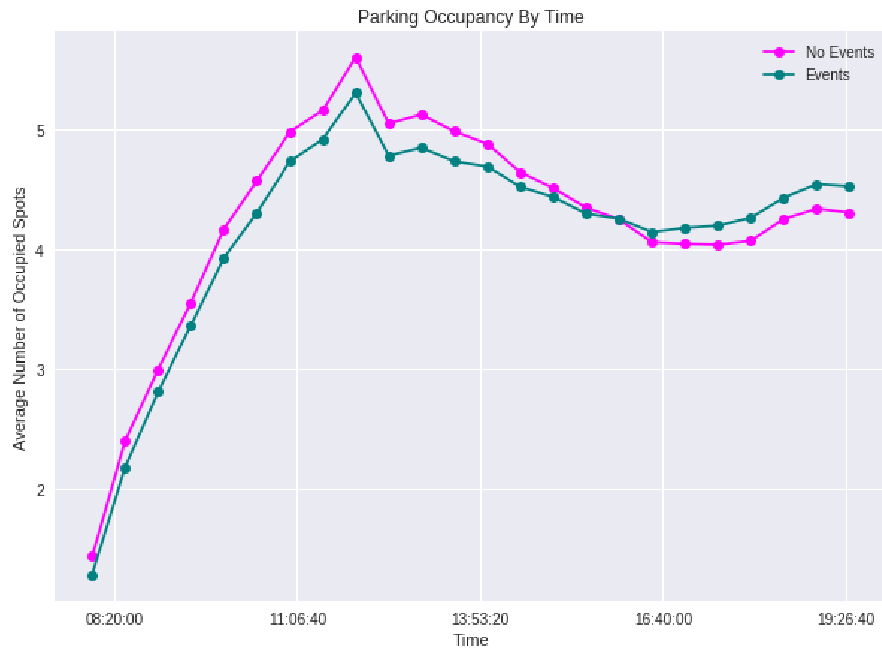


5. **Average** `occupied_spots` **(AUTOMATIC transactions) generally peaks in summer months and has a low point in December,** but variation is minimal**.**



6. **There is not a straightforward relationship between events and occupied spaces.** Events seem to have the most significant effect in the middle of the week, on Saturdays, and later in the day after 4:30pm. This might be related to event type and will be investigated further in modeling. This could also be due to data limitations, since event location is not exact.



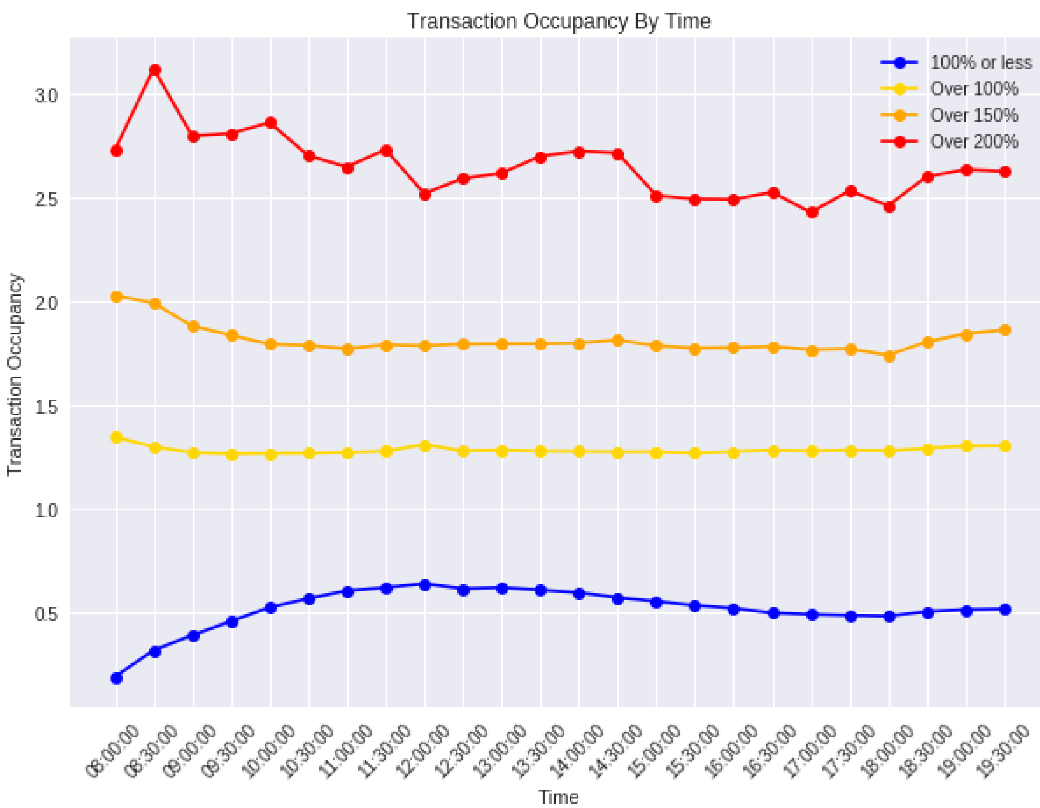


7. **8.1% of records have a** `transaction_occupancy` **value greater than 100%.** During EDA, the team added `transaction_occupancy` variable to reflect `occupied_spots` divided by `total_spots` for automatic transaction data. This rate could be due to overpayment (someone vacating a spot before their time runs out), paying at a kiosk not located on the blockface where they park, or system errors.
    a. Only 0.75% of records have a value for this field greater than 150%, and only 0.05% have a value greater than 200%. Based on the findings described below, the team will discuss how to handle pay stations or times of day at which errors occur more frequently.
    b. The team separated data into four categories for the next level of analysis: data with `transaction_occupancy` at 1) 100% or below, 2) >100%, 3) >150%, and 4) >200%.
    c. The team explored whether these errors were concentrated at specific `source_element_key` pay stations. After analysis, the team identified a few pay stations with higher error rates.
        i. For each potentially problematic pay station, what percent of their total records have `transaction_occupancy` greater than 100%?
            1. `source_element_key` 24041 : 47.96%
            2. `source_element_key` 81290 : 37.4%
            3. `source_element_key` 1025 : 35.71%
            4. `source_element_key` 31890 : 32.35%
            5. `source_element_key` 81294 : 31.35%
        ii. For each potentially problematic pay station, what percent of their total records have `transaction_occupancy` greater than 150%?
            1. `source_element_key` **1025 : 13.71%**
            2. `source_element_key` 27281 : 9.63%
            3. `source_element_key` 28953 : 6.57%

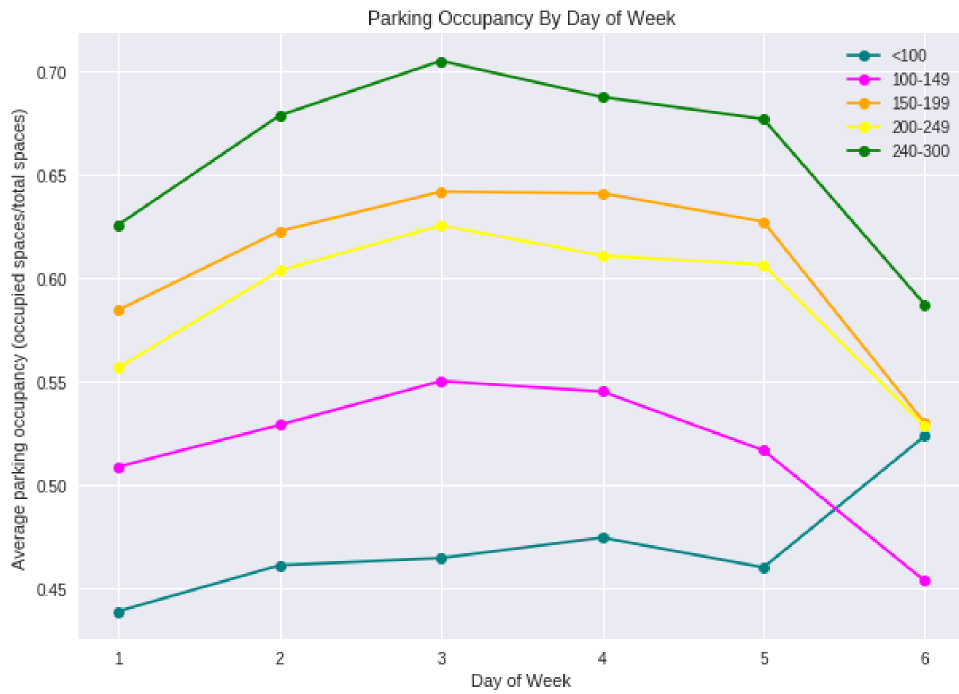
4. `source_element_key` 32434 : 6.61%
5. `source_element_key` 28962 : 5.73%

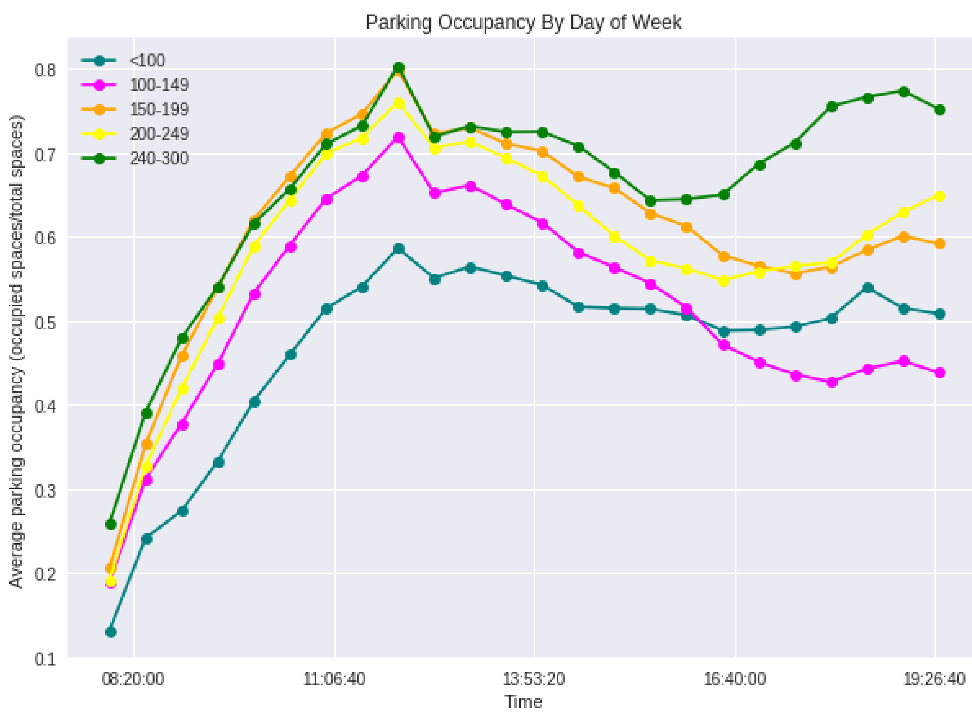
d. The team also explored whether error rates were concentrated across specific days of week, months, or times of day. No major trends were noticed across day of week or month. It's possible that `transaction_occupancy` rates tend to be slightly higher on Saturday for records with >100% values compared to 100% or below. So Saturday may observe slightly more human/system errors.

e. However, the primary finding here comes from time of day analysis. The blue line shows average `transaction_occupancy` values across the day for records with occupancy at 100% or below. This trend is similar to the graph on page 1, which makes sense. The difference appears when analyzing the middle two lines (gold and orange). **It appears that `transaction_occupancy` errors occur at higher rates at the beginning of the day and then decrease, which is the opposite of the blue line trend.** The red line is erratic, which could be due to the small sample size (only 1193 records).



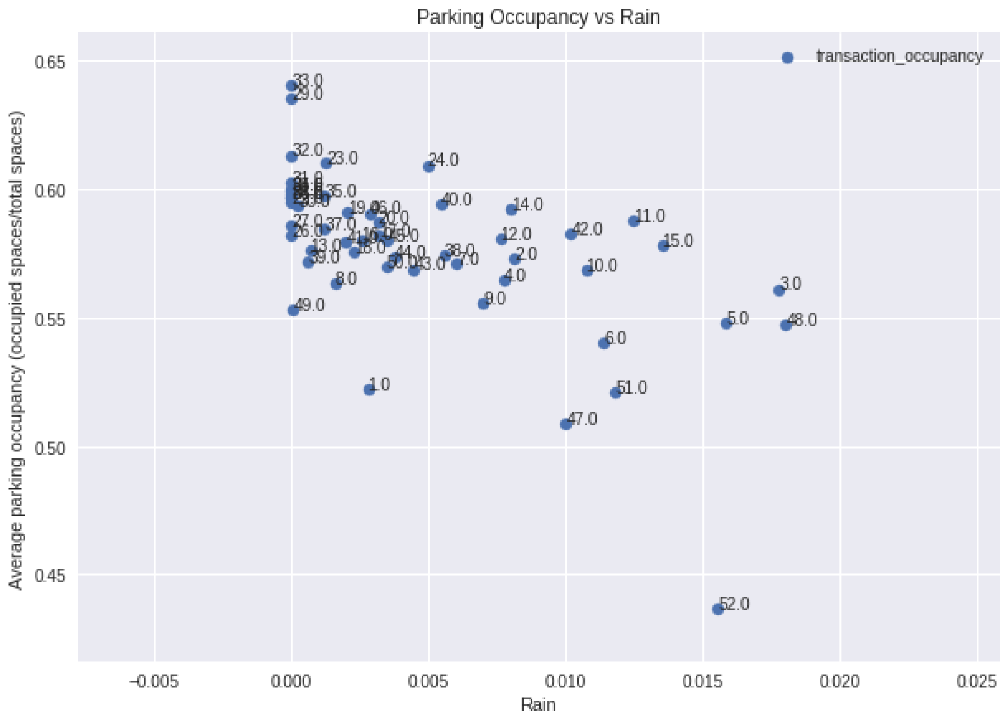
8. T**he graphs below show an apparent relationship between average parking occupancy and the number of nearby business as measured by business licenses.** During modeling, the team will validate this relationship and explore whether particular subcategories of licenses are more impactful.

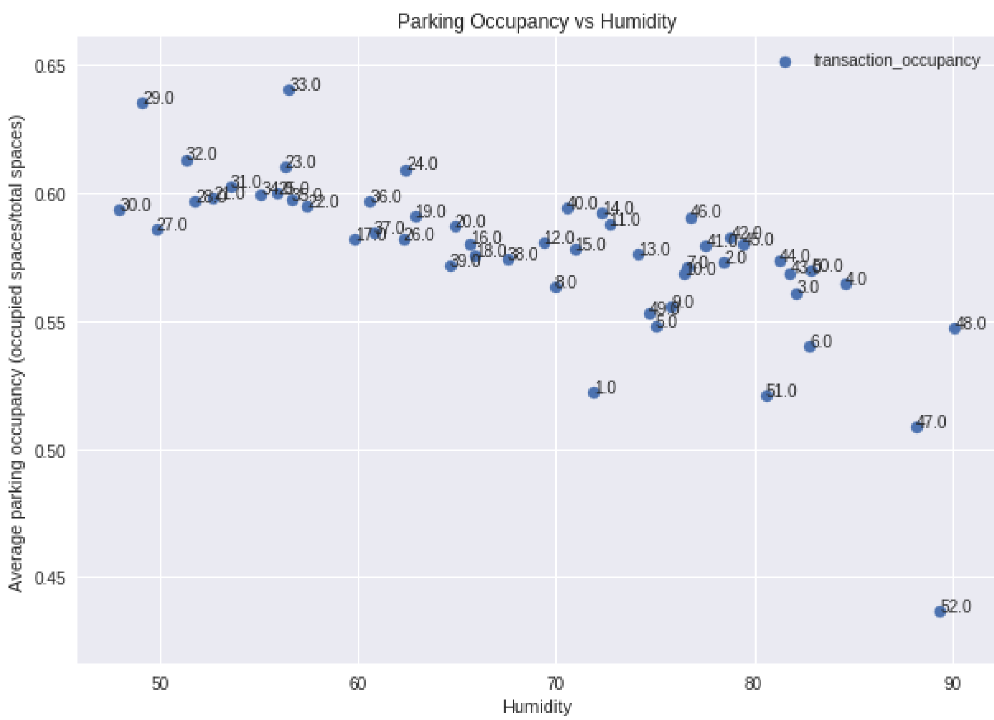




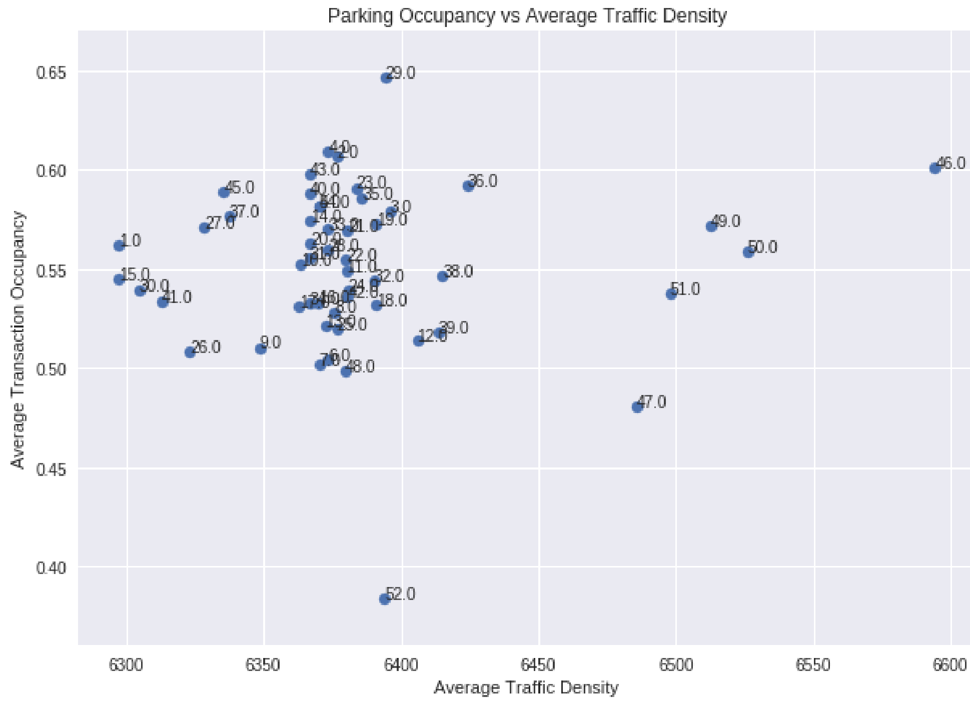
9. **It appears that** `transaction_occupancy` **tends to decrease as** `rain_inches` **increases.** `transaction_occupancy` also tends to decrease as `humidity` increases (humidity and rain are obviously related weather patterns). It is important to note that the team is not saying rain causes parking occupancy to decrease. Recall from above that parking occupancy also varied from month to month (higher in dry summer months and lower in wet winter months). In

the graphs below, the points are labeled with their corresponding `week_transaction` value (1 = first week of year, 52 = last week).





10. **There does not seem to be a relationship between** `transaction_occupancy` **and** `avg_traffic_density`**.** The latter variable takes data from the traffic flow volume datasets.



11. The parking occupancy from annual survey and transaction data follow a similar trend for the days the survey was conducted in 2017 and 2018. (Please note that `transaction_occupancy` > 100% has not been adjusted, and manual survey has not yet been scaled).