Predicting the likelihood of available parking spaces

Problem Definition

Vehicles spend more than 90% of their existence parked, which means finding an available parking spot is a challenge and takes a lot of time and patience to deal with. In this survey 61% of drivers said they feel stressed when looking for a parking spot, 42% said they missed appointments, 34% canceled trip because of parking challenges and 23% experienced road rage.

And it is not good for cities either. Searching for parking space can add up to 30% of the traffic in dense part of a town. This means a higher likelihood for accidents. And based on the INRIX study (2017), it costs the US economy $72.7 billion each year in wasted time, fuel and emissions. On average each driver spends 17 hours per year on searching for parking, which costs them $345 in wasted time and fuel.

This shows that searching for a parking spot has broader detrimental effects than just putting the driver in a bad mood. It affects other traffic participants, cities, countries, the economy and let's not forget about the environment.

Data Wrangling

The main aim of our project was to analyze and predict paid parking availability in the city of Seattle. We started by taking data for the same 30 day period from 2018 and 2019, available at the City of Seattle open data portal. The 30 day timeframe was from 21.03.2019 to 20.04.2019. While the data for 2019 was already available in the desired timeframe, the data for 2019 had to be extracted from a 50 GB dataset with 300MM entries.

To make our work with the datasets more computationally efficient we started with separating the 'OccupancyDateTime' timestamp into individual columns: year, month, day of the month, day of the week, time of day (hh:mm:ss), part of the day (AM or PM), hour of the day and minute of the hour. With the help of the split function we seperated Longitude and Latitude into individual columns from the 'Location' column.

We went on to add a 'Available_spaces' column, which is the subtraction of 'PaidOccupancy' column from the 'ParkingSpaceCount' column, to see how many parking spaces were available for each entry. Some values in this column were negative, because there were more vehicles parked in a location than the city estimated. We decided to add another column called 'AdjustedParkingAvailability' and changed those numbers to zero, to treat those parking location as fully occupied at that time.

The value in the ‘Zip Codes’ column were not actual Zip Codes. Fortunately each entry had a Longitude and Latitude column. With this information and the geolocation shape files with zip code polygon outlines from the King County website, we were able to find the corresponding zip code for each entry using Geopandas and Shapely libraries.

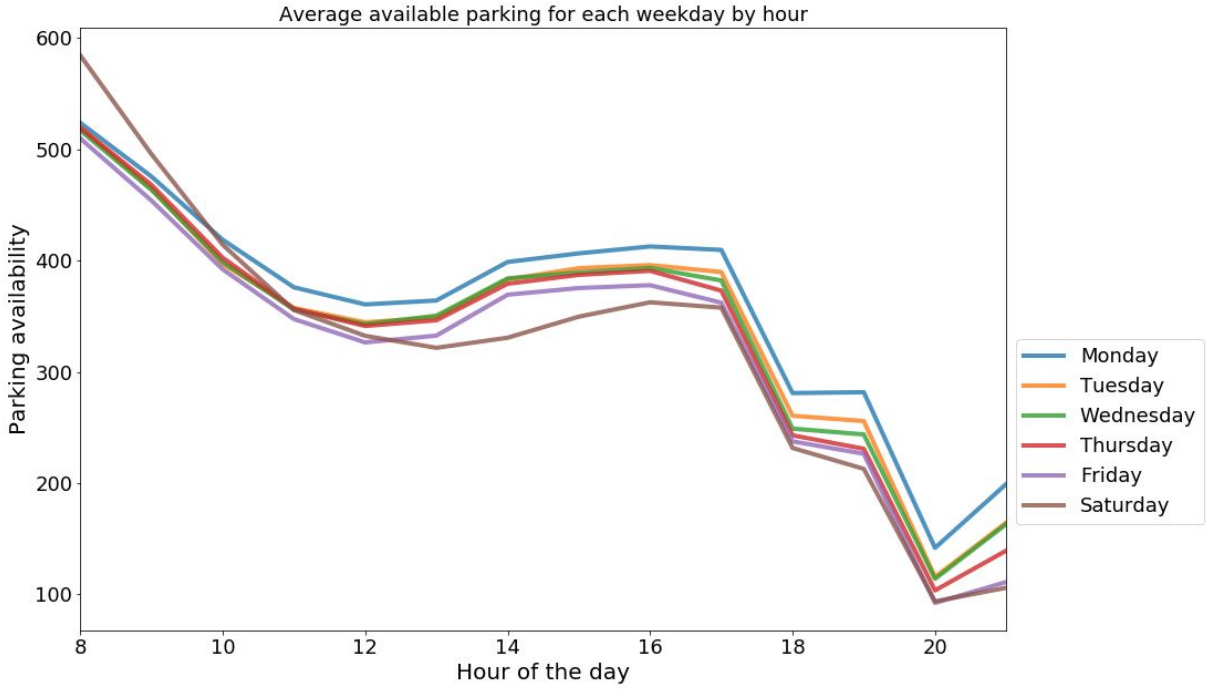
Paid Parking Availability in the city of Seattle

After getting more familiar with the dataset we formed some of the questions we are interested in answering with a deeper exploration:

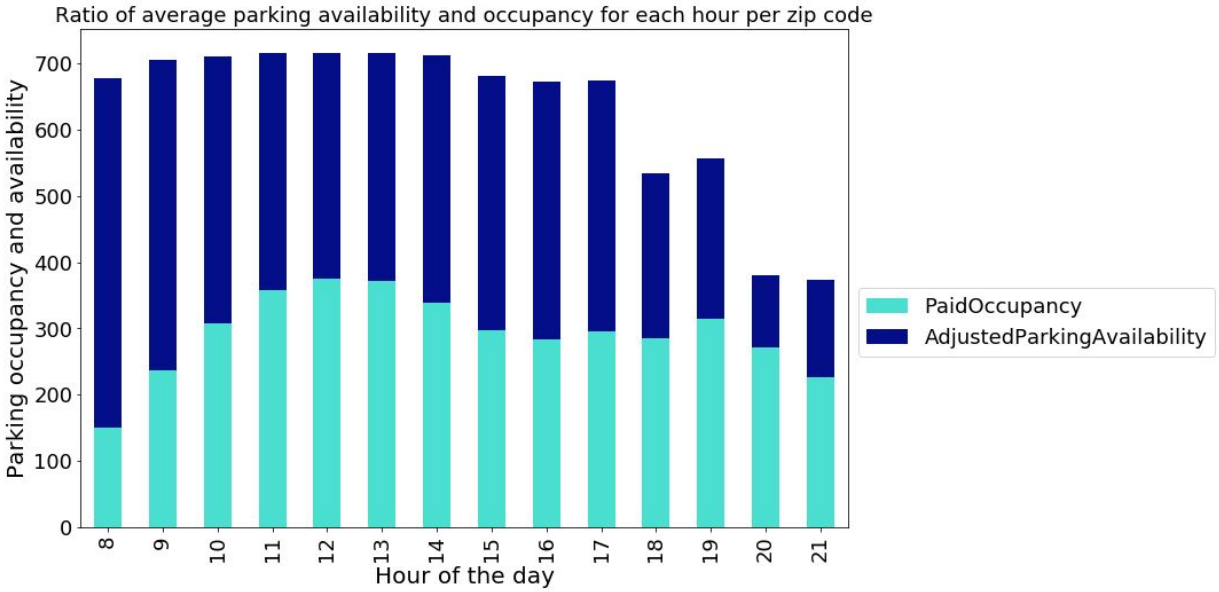
- How does time-of-day, day-of-the-week and location affect parking availability?
- Can we notice any patterns?
- How well does the city serve the supply and demand of paid parking? Is there enough parking available across the city?

Exploration Based on Time-of-day and Day-of-the-week

The graph below shows average parking availability across days of the week (excluding Sunday when parking is free) between 8:00 and 22:00 (for the hours outside of that parking is also free). From the graph we can see that for the most part the availability is similar across days of the week. The only outlier is Saturday, which is expected since most people don’t work on those days and are on a different schedule.

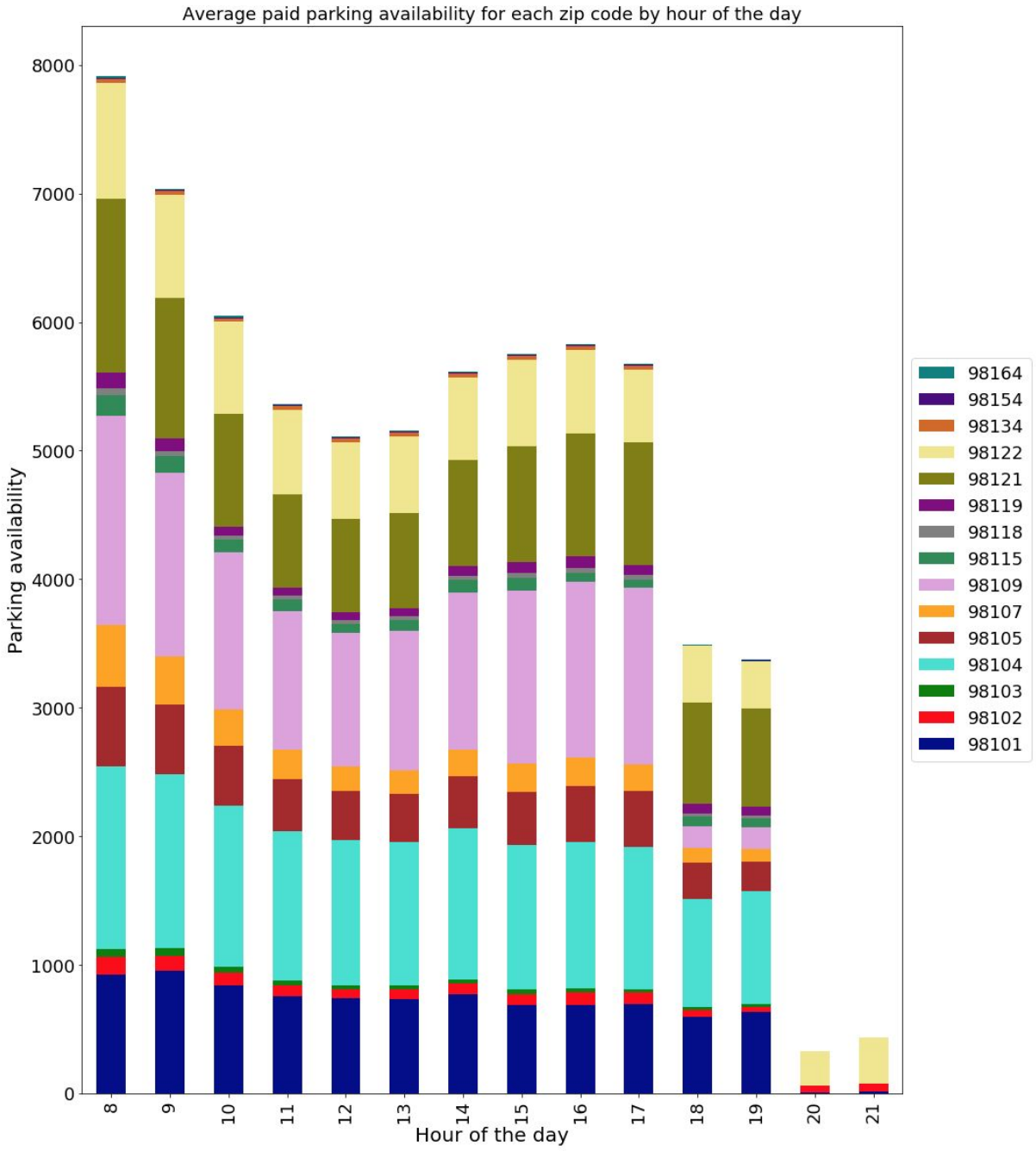


To make sure that occupancy and availability are as tightly connected as it appears in the graphs above, we created a stacked bar graph. The expectation would be that the top of the bar graph should be flat. This is true for the most part, except for two sudden drops one at 18:00 and another one at 20:00. The reason for those is that in certain zip codes parking becomes free after 18:00 or 20:00, which changes the average availability of paid parking across the zip codes.



Exploration Based on Location

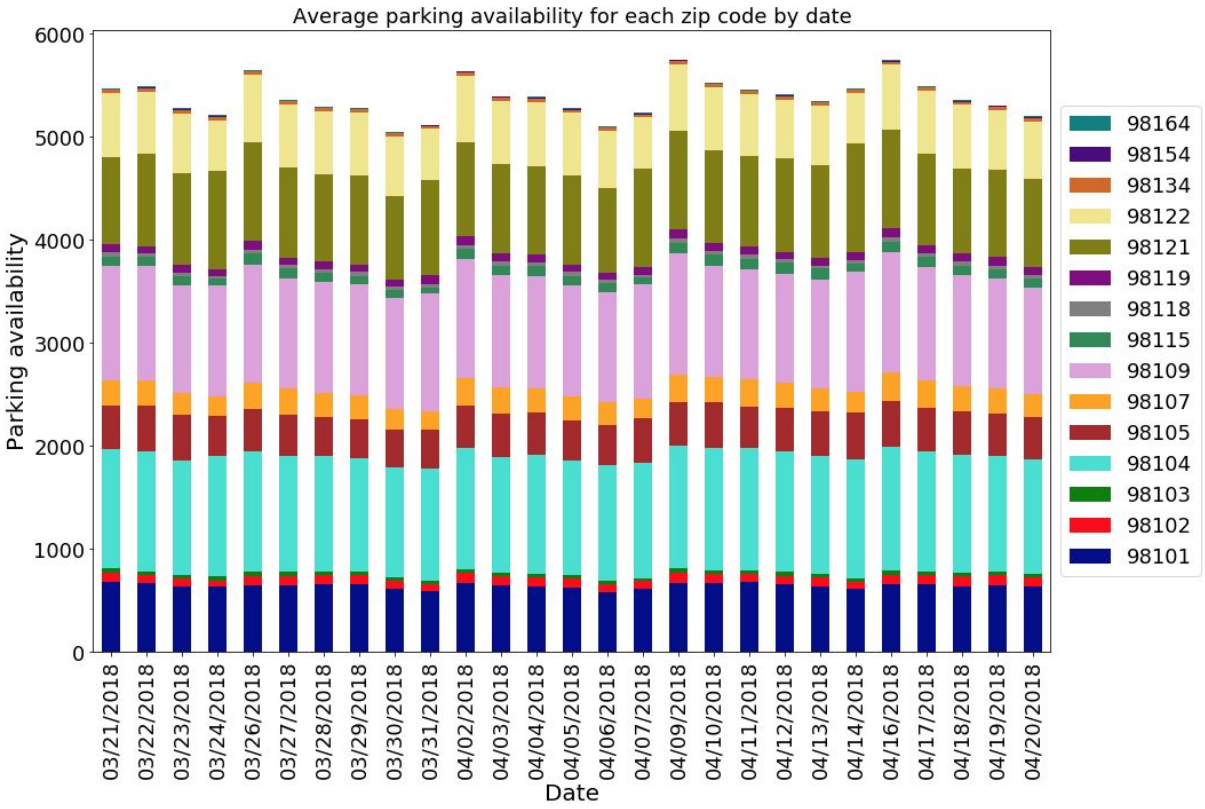
When we look at average parking availability based on location, in this case zip codes, we can notice a similar pattern emerge to the graphs above. But in this case we are able to identify how availability varies throughout the day for each zip code and identify the areas where paid parking ends sooner. Only three zip code locations - 98101, 98102, 98122 - have paid parking through most of the day, from 8:00 to 22:00. With the lowest availability of paid parking during the workday being from 12:00 to 14:00.



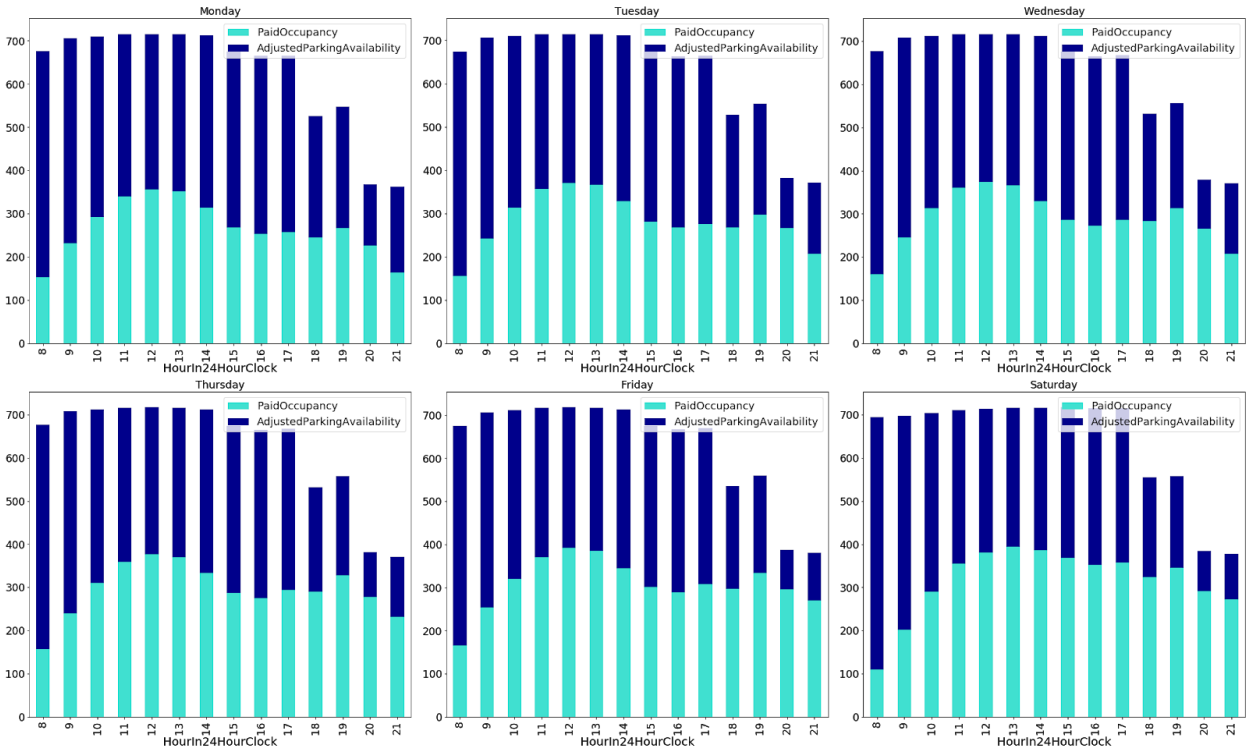
Patterns

So far we mainly looked at the data in aggregate across days of the week. If we plot out all the dates in our designated timeframe, we can notice a pattern emerge throughout most weeks,

where paid parking availability is highest at the beginning of the week and declines towards the end of it.

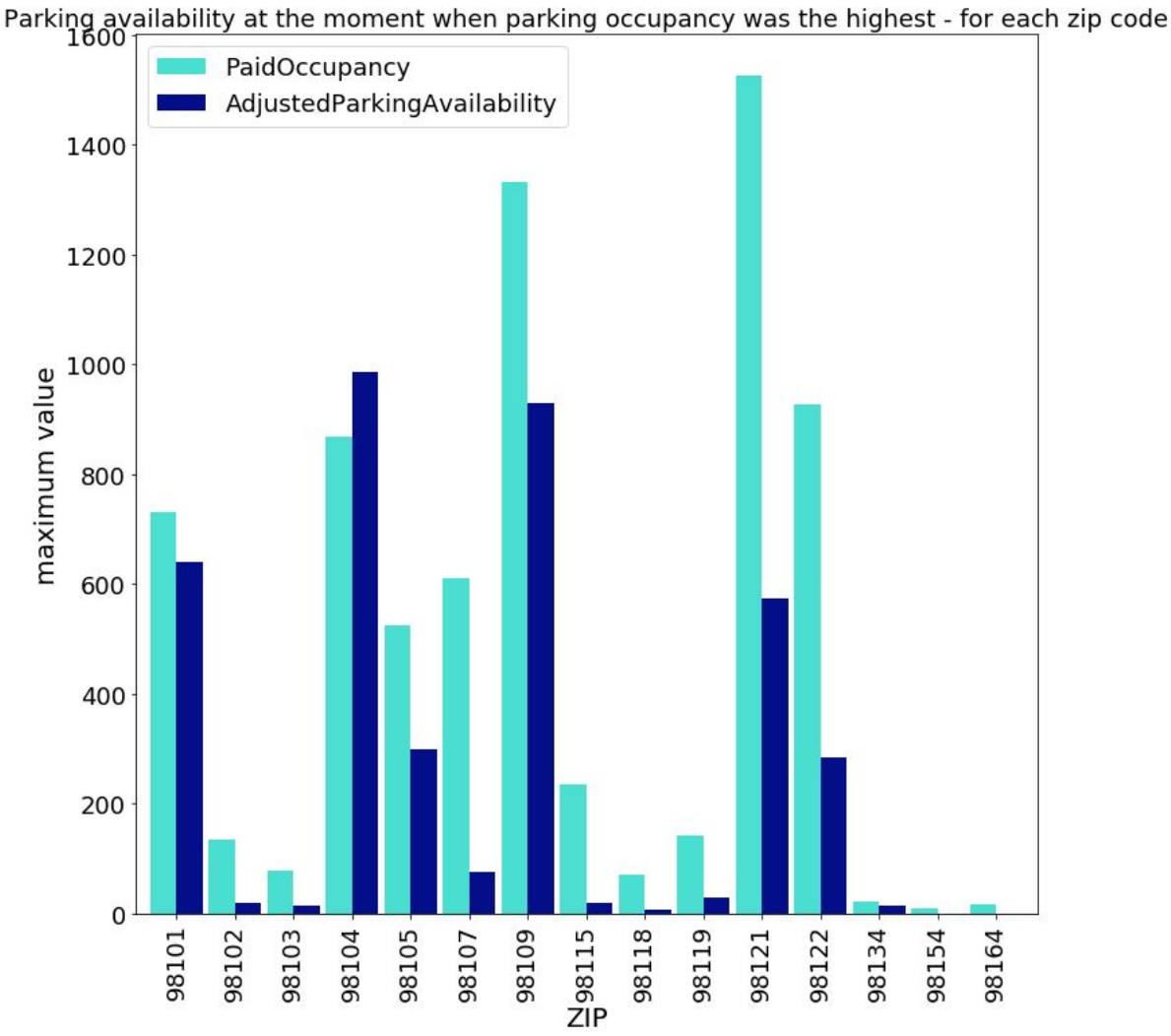


There are not only patterns on a weekly basis, but also in the hours of the day. At a glance there is very little variation between days of the week. The only exception being Saturdays.



Supply and Demand

We also wanted to understand how well the city is serving the needs of people in terms of paid parking availability in Seattle. We looked at one occasion for each zip code where parking occupancy was at its peak and looked at the availability for that moment in time. Especially areas with generally more availability still had space available event at its peak. The exception being 98107, which also includes Ballard Locks. Zip codes with less general availability were close to or at capacity at the time of their peak.



Data Story Conclusion

We are able to establish that paid parking availability is fairly predictable for the city of Seattle and the explored time frame from 21.03.2018 to 20.04.2019. The data follows weekly as well as daily and hourly patterns. There were very few outliers or anomalies. What that tells us is that people follow predictable day-to-day patterns when it comes to parking. And apart from a few areas with a smaller number of paid parking spaces available, the people of Seattle seem to have ample paid parking options. At least based on the data at hand.

Inferential Statistics

For our statistical analysis we decided to use the one-way ANOVA (analysis of variance), to compare the means of groups by analyzing variances. We have 6 days of the week (Monday,

Tuesday, Wednesday, Thursday, Friday and Saturday), which represents groups in our ANOVA analysis. Days of the week are independent variable. Data for parking availability presents dependable variable.

ANOVA hypothesis:

null hypothesis: There is no difference in means of parking availability between days of the week (excluded Sunday).

alternative hypothesis: There is a difference between the means of parking availability between days of the week (excluded Sunday).

ANOVA table with model effect size

The effect size tells us how much of an impact the experiment will have in the real world. There are a few different effect sizes one can use: eta squared (eta_sq), and omega squared (omega_sq). Omega squared is considered a better measure of effect size than eta squared because it is unbiased in it's calculation.

	sum_sq	df	mean_sq	F	PR(>F)	eta_sq	omega_sq
C(Weekday)	4.822448e+05	5.0	96448.954315	6680.720793	0.0	0.0013	0.001299
Residual	3.706132e+08	25671230.0	14.436908	NaN	NaN	NaN	NaN

The Weekday row is the between groups effect which is the overall experimental effect. The sum of squares for the model (sum_sq = 4.822448e+05) is how much variance is explained by the model. The current model explains a significant amount of variance, $F(5,25671230)= 6681$, $p < 0.05$.

The Residual row is the unexplained variance in the data (sum_sq = 3.706132e+08). In this case, the unexplained variance represents the individual differences in parking availability and different reactions to a day of the week.

The mean_sq columns tell us, the average amount of explained variance (96448.95431536864) and the average amount of unexplained variance (14.436908426270824) by the current model.

The F value is the point such that the area of the curve past that point to the tail is just the p-value, therefore PR(>F)=p. Based on the p-value (0.0), we can say there is a significant difference in the group means.

The column eta_sq tell us, that the current model accounts for 0.13% of the variance in contributing to parking availability.

Other tests

We used Levene's test to check for homogeneity of variance. As the p-value (0.0) is significant, which indicates that the groups don't have equal variances.

We also checked the normal distribution of residuals**.** The results from the Shapiro-Wilk test is statistically significant, which indicates that the residuals aren't normally distributed.

Post-hoc Testing

We used Tukey HSD post-hoc comparison test between different group means. And the results are in the following table.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Friday	Monday	0.4074	0.001	0.3999	0.4148	True
Friday	Saturday	-0.0159	0.001	-0.0233	-0.0084	True
Friday	Thursday	0.126	0.001	0.1189	0.133	True
Friday	Tuesday	0.2153	0.001	0.2079	0.2228	True
Friday	Wednesday	0.1635	0.001	0.1564	0.1705	True
Monday	Saturday	-0.4232	0.001	-0.431	-0.4154	True
Monday	Thursday	-0.2814	0.001	-0.2889	-0.2739	True
Monday	Tuesday	-0.192	0.001	-0.1999	-0.1842	True
Monday	Wednesday	-0.2439	0.001	-0.2513	-0.2364	True
Saturday	Thursday	0.1418	0.001	0.1344	0.1492	True
Saturday	Tuesday	0.2312	0.001	0.2233	0.239	True
Saturday	Wednesday	0.1793	0.001	0.1719	0.1868	True
Thursday	Tuesday	0.0894	0.001	0.0819	0.0968	True
Thursday	Wednesday	0.0375	0.001	0.0305	0.0445	True
Tuesday	Wednesday	-0.0519	0.001	-0.0593	-0.0444	True

The Tukey HSD post-hoc comparison test controls for type I error and maintains the familywise error rate at 0.05 (FWER= 0.05 top of the table). The group1 and group2 columns are the groups being compared. The meandiff column is the difference in means of the two groups (group2 – group1). The lower/upper columns are the lower/upper boundaries of the 95% confidence interval. The reject column tells us that we should reject the null hypothesis.

We can conclude that there are statistically significant differences between groups (days of the week).

Machine learning

Techniques used

For this project we decided to use the Linear Regression and the Random Forest method. We tried Linear Regression, because it is appropriate to use if a dependable variable (in our case Parking availability) has a linear dependency with the predictors (independent variables). This technique also enables an estimate of quantity (number of available parking spots). In case of more complex dependencies (when there is no linear dependence), using Random Forest is more appropriate. Random Forest also enables us to determine which class the dependant variable (parking availability) belongs to (ZIP code, parking area, etc). This project is far more detailed than it is outlined in this report. For a deeper analysis with all the models and graphs, check out the files on Github.

Results for each technique

For both methods we tried different approaches to test which ones provide the best results. We had the same data grouped in two different ways by area, one based on ZIP codes and one based on city defined parking areas.

Linear Regression using statsmodels

We got better results and better fit for data when using three categorical features instead of just one, or a set of two. These three features were *Day of the week* and *Hour of the day*, *ZIP codes* or *Parking area*, but (surprisingly) not Parking Occupancy.

Performance

The model for data grouped by ZIP code, had an R_quared of 0.905, F-statistic of 88640 and p-value of 0. These tell us that the model is statistically significant, i.e. the predictors are jointly informative. And that 90.5% of the variance in the response variable (Parking availability) can be explained by this model. All variables had p-value less than zero, that indicates they are significant and likely to influence parking availability.

The model where data were grouped by parking area, had an R_squared of 0.312, F of 7749, and p-value of 0, which means that overall the model was statistically significant. But in this model there were 3 parking areas that weren't significant, Columbia City, Fremont and Green Lake. These three areas are less likely to impact predicted parking availability.

In both models parking availability decreases between 11am and 13pm, with a peak at 12pm. Another similarity is that Saturdays have the least parking availability and Mondays have the most.

Even though the model based on ZIP codes had the largest R_squared, which means that those models describe more variance than any other model based on parking area. Based on these models we cannot assume that ZIP codes in the city center influence the parking availability more than ZIP codes on the

outskirts. On the other hand, this doesn’t apply to parking areas. As we already mentioned there are three parking areas that aren’t that influential on predictions and they are not in the city center.

For this reason we also built models for specific parking areas to see how they perform. We chose South Lake Union, Belltown and Columbia City. One thing to note, Belltown is in the city center, South Lake Union is next to the city center and Columbia City is south from the city center. Overall each model was statistically significant, but R_squared was low. By surprise Columbia City had the largest R_squared, which indicated that Weekday and HourIn24HourClock variables can explain more parking availability in that parking area than in any of the other two. Columbia City has the lowest estimated average parking availability when our independent variables (Weekday and HourIn24HourClock) are zero. In South Lake Union and Belltown, Tuesday, Wednesday and Thursday have a p-value greater than 0.05, which means they aren't statistically significant and because of that less likely to influence parking availability in that parking area.

Based on this information we can only assume that in the city center Tuesday, Wednesday and Thursday don't affect parking availability as much as other days of the week do. We can assume that regardless of parking area, parking availability decreases around 12pm. Drivers are less likely to find a parking spot around that time.

OLS					
Data	MODEL	R_Squared	F-statistic	p-value	AIC
OccupancyDateTime	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.037	54640	0	140500000
Parking areas	AdjustedParkingAvailability ~ PaidOccupancy'	0.481	588000	0	8217000
ZIP codes	AdjustedParkingAvailability ~ C(ZIP)'	0.849	119600	0	3896000
ZIP codes	AdjustedParkingAvailability ~ C(ZIP) + C(Weekday) + HourIn24HourClock'	0.884	114000	0	3817000
ZIP codes	AdjustedParkingAvailability ~ C(ZIP) + C(Weekday) + C(HourIn24HourClock)'	0.905	88640	0	3758000
ZIP 98101	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.664	7471	0	295300
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.978	55840	0	233500
ZIP 98121	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.225	941.7	0	255900
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.832	6019	0	226200
ZIP 98115	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.492	3140	0	177000
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.77	4068	0	161600
	AdjustedParkingAvailability ~ C(PaidParkingArea)'	0.295	13950	0	8412000
	AdjustedParkingAvailability ~ C(PaidParkingArea) + C(Weekday) + HourIn24HourClock'	0.304	11090	0	8403000
	AdjustedParkingAvailability ~ C(PaidParkingArea) + C(Weekday) + C(HourIn24HourClock)'	0.312	7749	0	8397000
South Lake Union	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.024	125.6	1.26E-157	474300
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.035	75.71	2.05E-228	473900
Belltown	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.006	58.01	6.68E-72	849500
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.022	79.5	4.51E-258	848600
Columbia City	AdjustedParkingAvailability ~ C(Weekday) + HourIn24HourClock'	0.393	2096	0	139100
	AdjustedParkingAvailability ~ C(Weekday) + C(HourIn24HourClock)'	0.747	3583	0	122100

Random Forest

When using the Random Forest method we did include Parking Occupancy, because as expected it is the most important feature to make predictions. Each random forest model had 100 estimators, except for random search which had 20 estimators.

Performance

To see how each model performed, we can look at its error sizes. Particularly max error which is the largest prediction mistake the model made. The Mean Absolute Error (or MAE) is the average of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were, in other words the magnitude of the error, but doesn't say anything about the direction (e.g. over or under predicting). The R_squared metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determination. This is a value between 0 and 1, for no-fit and perfect fit, respectively. To know about the accuracy of the model we can look at RMSE. Which is taking the mean of all prediction errors, squaring them, and finally taking the root will give us the RMSE of a model. The lower the value, the better the fit of the model to the data. The last column in the table tells us if the model is accurate or not.

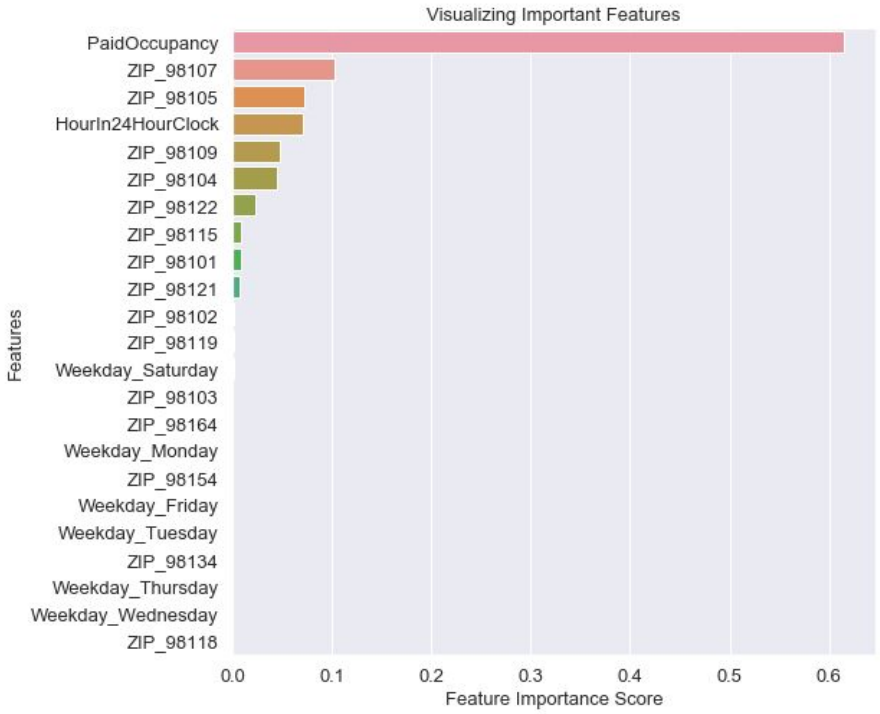
Random forest						
Model	Data	Max error	Mean absolute error	R_squared	Root mean squared error	Accuracy (RMSE < 10% of the mean value)
ZIP w Parking Occupancy	2018	165.89	4.45	0.9995	9.53	true
	2019	1728.7	44.02	0.9263	130.91	false
ZIP without Parking Occupancy	2018	249.76	17.3	0.9954	29.08	true
ZIP 98101	2018	101.98	10.87	0.9961	17.38	true
Parking areas without ZIP codes	2018	318.08	5.76	0.9933	18.06	false
	2019	459.12	20.41	0.9678	42.89	false
Parking areas with ZIP code	2018	99.66	2.82	0.9992	6.38	true
	2019	453.94	17.72	0.9730	39.24	false
Parking area with ZIP codes and without parking occupancy	2018	1007.16	10.33	0.9935	17.78	false
South Lake Union with ZIP code and parking occupancy	2018	92.30	8.94	0.9988	16.92	true
South Lake Union with ZIP code and without parking occupancy	2018	117.85	16.11	0.9971	26.15	true
Random search (parking areas with ZIP codes and parking occupancy)	2018	97.02	2.88	0.9992	6.12	true
Grid search and cross validation (parking areas with ZIP codes and parking occupancy)	2018	108.19	2.79	0.9992	6.33	true

By comparing max error and mean absolute error of the parking area model to the one with ZIP codes, the ZIP code model did a better job at predictions. For 2018 data the parking area model had a max prediction error of 318 while ZIP codes had 166 parking spots. The parking area model had a mean absolute error of 5.76 compared to the ZIP code model where it was 4.45 parking spots. Based on the value of the root mean squared error, which for ZIP code is lower than 10% of the mean value for parking availability (mean is 364.84), it means the model is accurate. But for the parking area model it is slightly greater than 10% of the mean value of parking availability (171.69), which means the model is not accurate. Looking at these error values, these two models cannot make reasonably good predictions. We also made predictions for 2019 data, based on a 2018 training set for the same time period. Here the results are a bit different and the parking area model does a better job with a max error of 459 and mean absolute error of 20.41, compared to 1729 and 44.02 for ZIP codes model. It seems that changes which occurred over the span of a year affected ZIP codes more than parking areas.

Going through all these models, we can see that the Random search model (marked green in the table), which was built with data based on parking areas with zip codes and parking occupancy, has the lowest errors. And the model ZIP code with parking occupancy has the largest errors and the lowest R_squared (marked red in the table above).

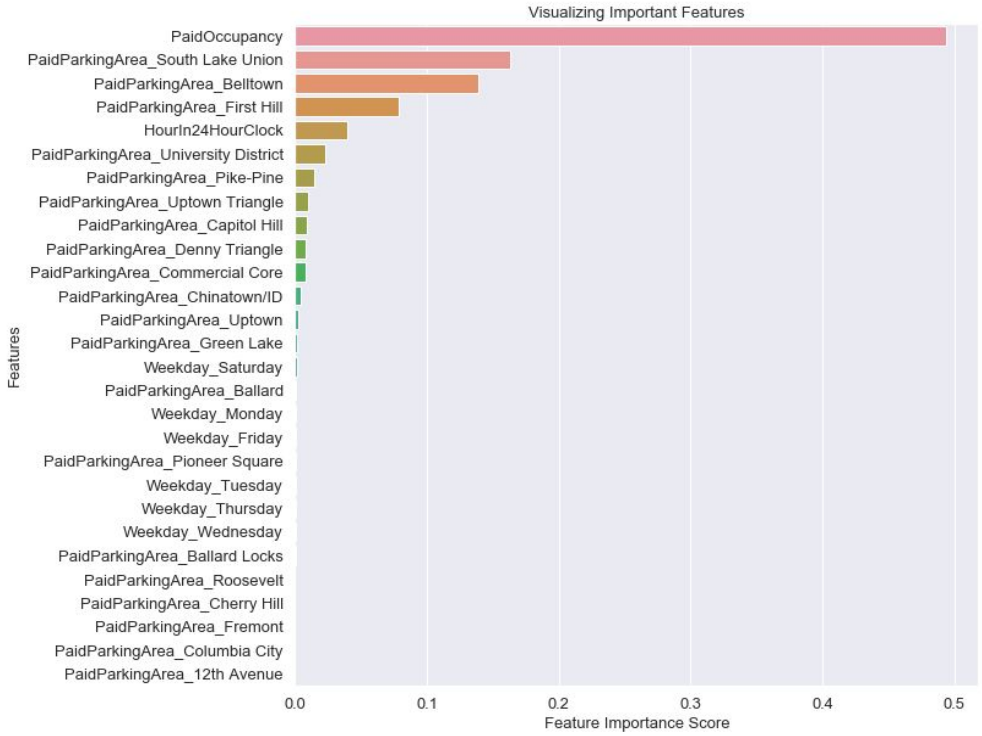
Feature importance

Now let's take a look at which features were the most important in making predictions. Feature importance represents how much including a particular variable into the model improves the prediction.



We got some unexpected results. The most important feature was Parking Occupancy, this was expected since it has the biggest correlation with parking availability. But the rest features varies between different models.

For models based on ZIP codes, the most important ZIP codes (on the second and third place) are not in the city center, but in the northern part. Following that is the hour of the day, which is a more important feature than the day of the week. This makes sense, since parking fluctuates throughout the day more than it does between days. At the end of the list is a ZIP code that is south of the city center. We cannot make assumptions that the size or location of the ZIP code plays a significant part in the order. For instance, ZIP code 98107 is smaller than 98118 and neither of them is in the city center.



For the parking area model, the order is different. After Parking Occupancy, follow South Lake Union (close to the city center) and Belltown which is in the city center. As expected Hour of the day is a much more important feature than the day of the week. Based on this order we can make the assumption that the location doesn't matter, but the size of a parking area does. For instance, Pioneer Square and Denny Triangle are in the city center but are not in the top 10 most important features. All top three parking areas are large compared to others and all the ones at the bottom are the smallest.

Conclusion

Random Forest works much better in our case compared to Linear Regression. When we compare true and predicted values, we see a better performance (lower prediction errors) with the Random Forest models. For this particular case, using Linear Regression was not the best option, because the relationships between variables don't seem to be behaving in a linear fashion.

By comparing the performance of different random forest models we built, we noticed that models in which the data is grouped by ZIP codes, or where ZIP codes are additional independent categorical variables, perform better. Those models had lower prediction errors than models for parking areas without zip codes as independent variable, even when making predictions on new, unseen data, for 2019. Overall, models for both ZIP codes and parking areas didn't perform well on new data (2019) compare to the training data (2018). This might probably be because of minor changes in the paid parking setup over the course of a year, and a possible shift in the behaviours of people.

In terms of feature importance, we expected hour of the day to be the top feature. It turns out the top predictor is *Parking Occupancy*, which makes sense because it's very closely linked to parking availability. But hour of the day came in 4th for ZIP codes and 5th for parking areas. The models put certain ZIP codes or parking areas ahead of the hour of the day.

In the city center, Belltown is a parking area with the highest amount of paid parking options. Right next to Belltown is South Lake Union, which has the highest estimated number of paid parking availability, on average. In Belltown and South Lake Union parking areas it might be challenging to find paid parking on Friday around 12pm. In both parking areas after 3pm it is much easier to find a paid parking spot compared to earlier hours of the day. And looking for available paid parking on Saturday doesn't seem to be a problem.

The most challenging to find available paid parking spots might be in Columbia City, which has on average the lowest estimated parking availability. Based on our model for the Columbia City parking area, it is not advisable to look for paid parking during Saturdays, especially not around 12pm. But between 2pm and 6pm it is more likely to find a spot. Overall the parking availability is decreasing constantly through the day. Since there aren't many paid parking garages or lots in that area, it's better to look for on-street parking.