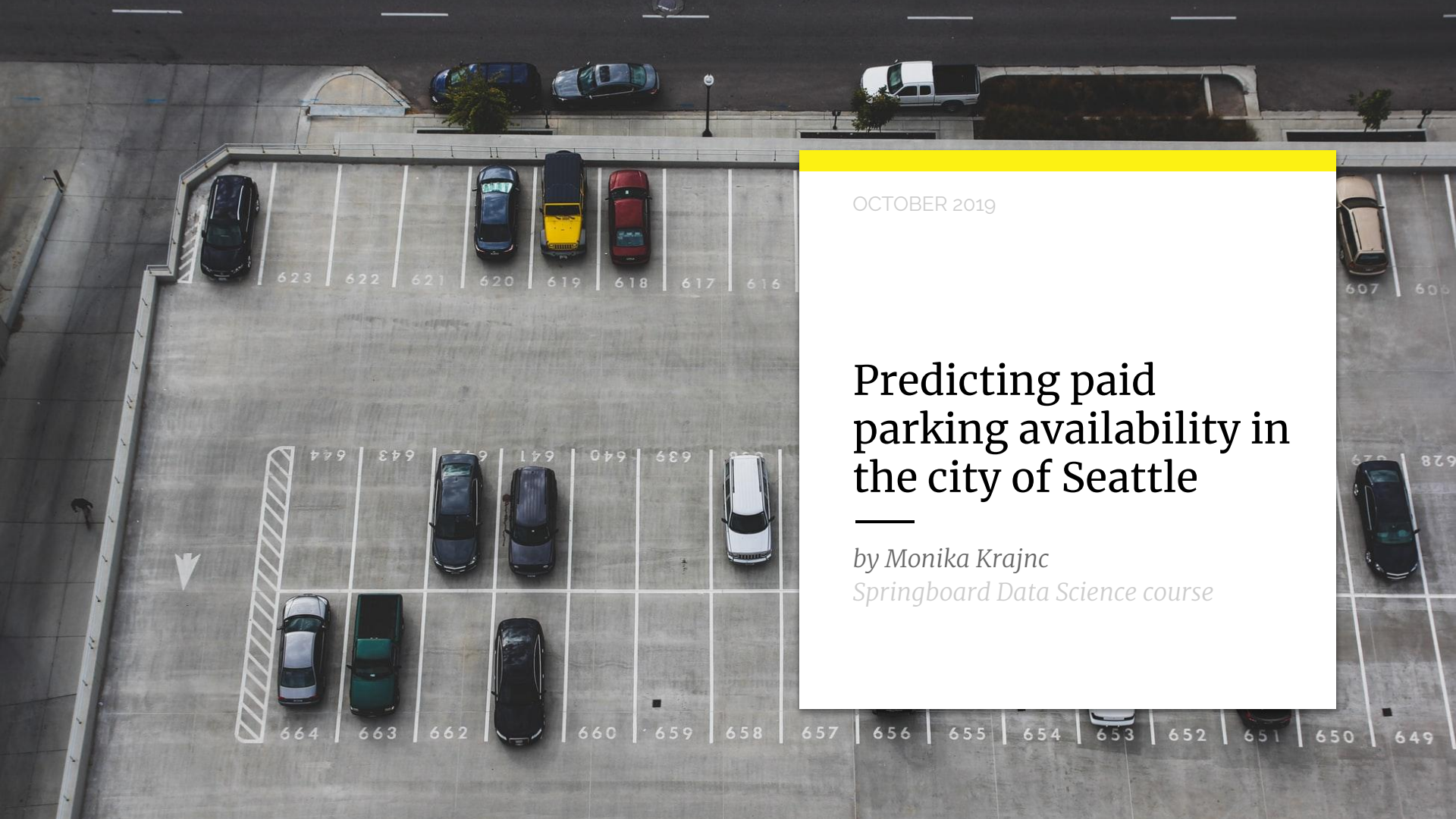


OCTOBER 2019

Predicting paid parking availability in the city of Seattle

by Monika Krajnc
Springboard Data Science course

Overview

The Main Question

While we live in a world of abundant data, live data about available parking spots is still hard to come by.

Instead of relying on live data, would it be possible to predict the likelihood of finding a parking spot in an area, before even driving there?

The Problem for Drivers

Finding parking is one of the major stressors and causes of emotional discomfort for drivers. In this survey ***61%*** *of drivers reported* ***feeling stressed*** *when looking for a parking spot,* ***42%*** *said they* ***missed appointments***, ***34% canceled a trip*** *because of parking challenges and* ***23% experienced road rage****.*

The Problem for Cities and the Economy

Looking for parking is not good for cities or the economy either.

*This article reports that searching for a parking space can represent up to **30% of the traffic** in dense part of a town. In addition to that, based on a recent INRIX study, searching for a parking space costs the US economy **$72.7 billion each year** in wasted time, fuel and emissions.*

The Business Opportunity

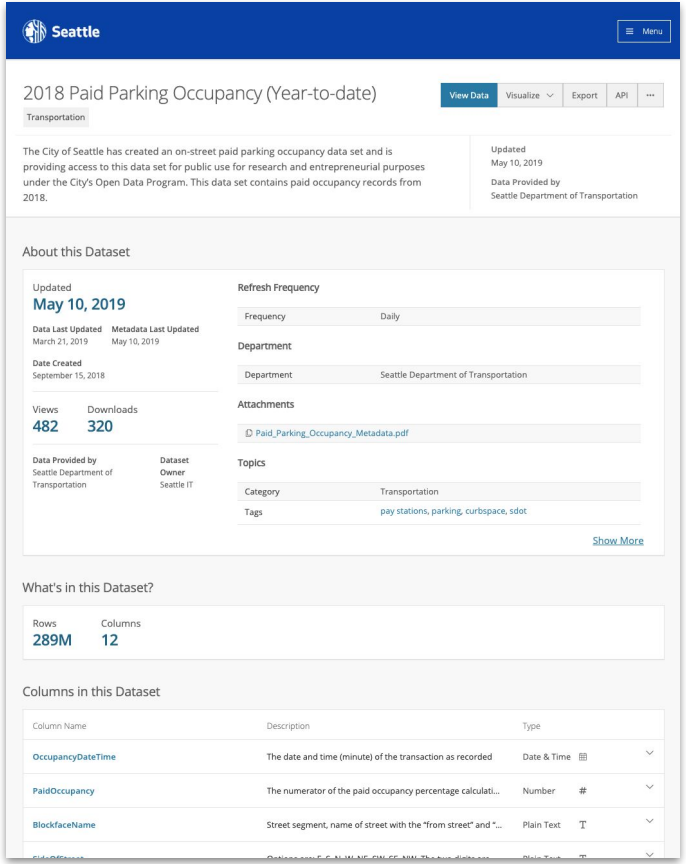
Navigation providers like Garmin, TomTom, HERE or even Waze, as well as automotive companies could offer parking availability prediction features to customers. The demand is already there since ***72% of respondes*** *of a recent INRIX survey would love to have a feature like that.*

The Business Opportunity

Local shops and business are also affected by unpredictable parking availability. In this survey 63% of the drivers responded that they avoid driving to specific destination if it's hard to find parking space there. And 39% of those drivers are also avoiding shopping in such areas. Leading to potential lost sales for business in areas with less accessible parking.

The Data

For this project we decided to use paid parking data from the City of Seattle, since the dataset was very extensive and well structured. We used parking data from 2018 and 2019, as well as zip code boundary data. Data for free parking is unfortunately not available.



Exploratory Analysis and Findings

Data Wrangling and Preparation

- *Extracted a month's worth of data from the 2018 dataset which was 50GB and contained close to 300M rows.*

- *Separated the timestamp column into individual components for computational efficiency, into 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.*

- *More details about this process can be found here*

Data Wrangling and Preparation

- *Calculated available spaces by subtracting the 'PaidOccupancy' from the 'ParkingSpaceCount' column. There were situations with more occupancy than availability, which could be attributed to more vehicles occupying a designated area. We treated every situation exceeding 'ParkingSpaceCount' as zero parking available.*

- *Rearranged the data from being grouped by parking area to zip codes, by using coordinates in the data and zip code polygon outlines.*

Our Questions

- *How does time-of-day, day-of-the-week and location affect parking availability?*

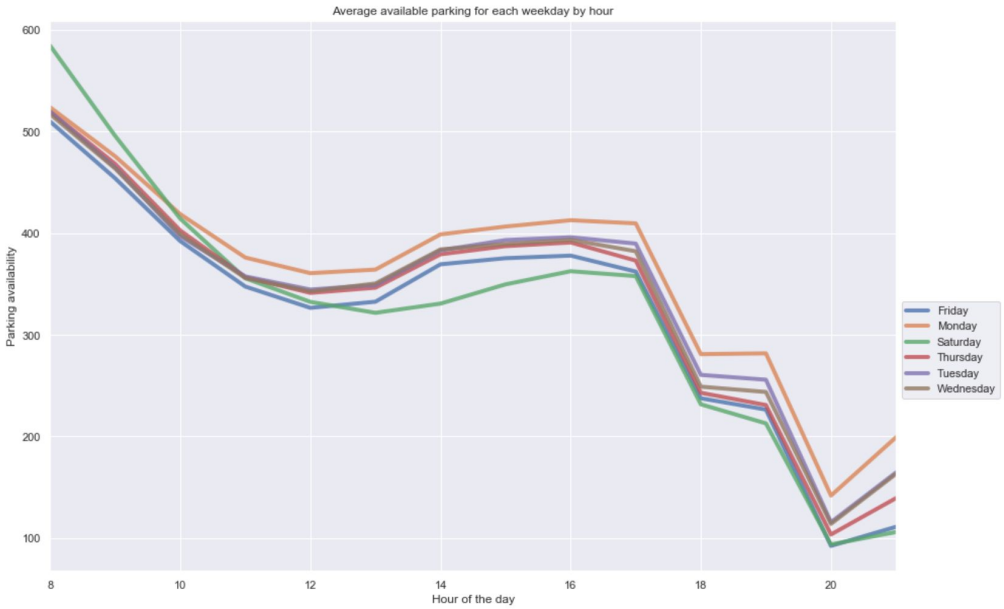
- *Can we notice any patterns?*

- *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

This graph shows average parking availability across days of the week between 8:00 and 22:00 (for the hours outside of that, as well as Sundays parking is free of charge).

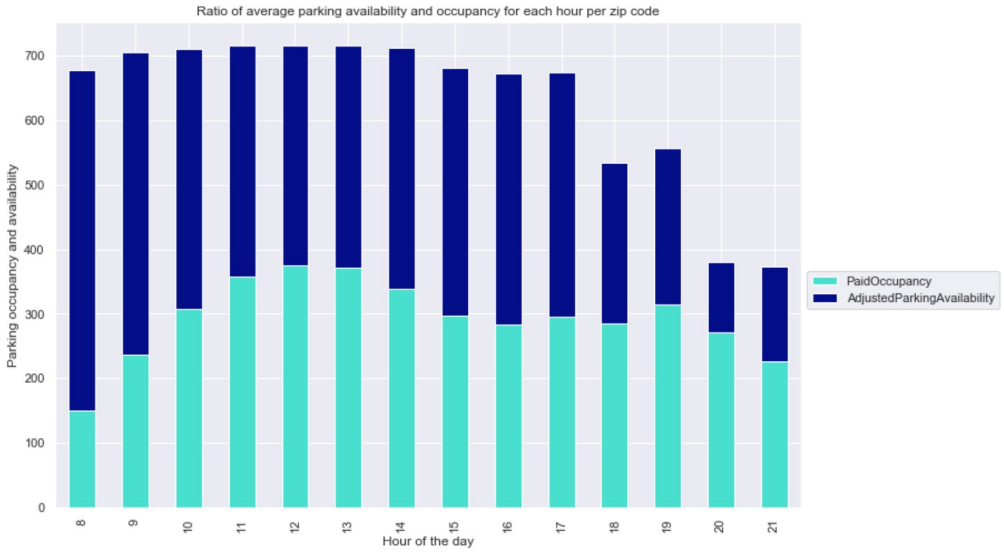
From the graph we can see that for the most part the availability is similar across days of the week. The only outlier is Saturday, which is expected since most people don't work on those days and are on a different schedule.



Exploration Based on Time

The expectation on this stacked graph of availability and occupancy would be that the top of the bar graph should be flat.

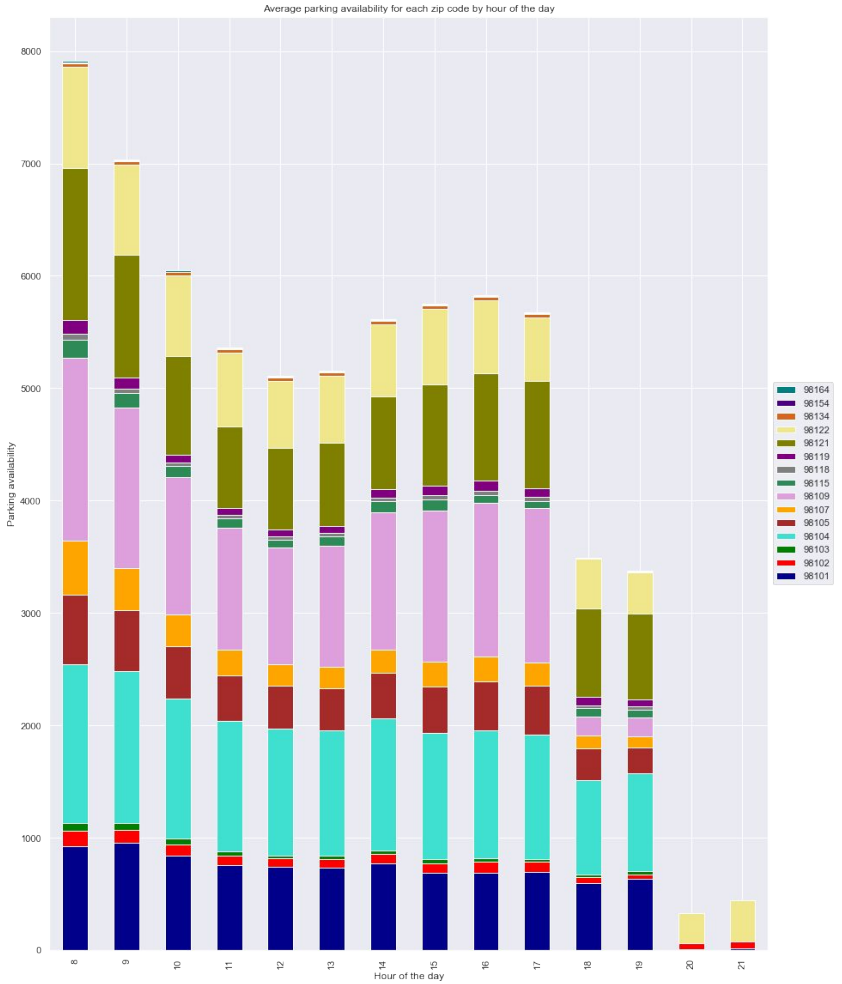
This is true for the most part, except for two sudden drops one at 18:00 and another one at 20:00. The reason for those is that in certain zip codes parking becomes free after 18:00 or 20:00, which changes the average availability of paid parking across the zip codes.



Exploration Based on Location

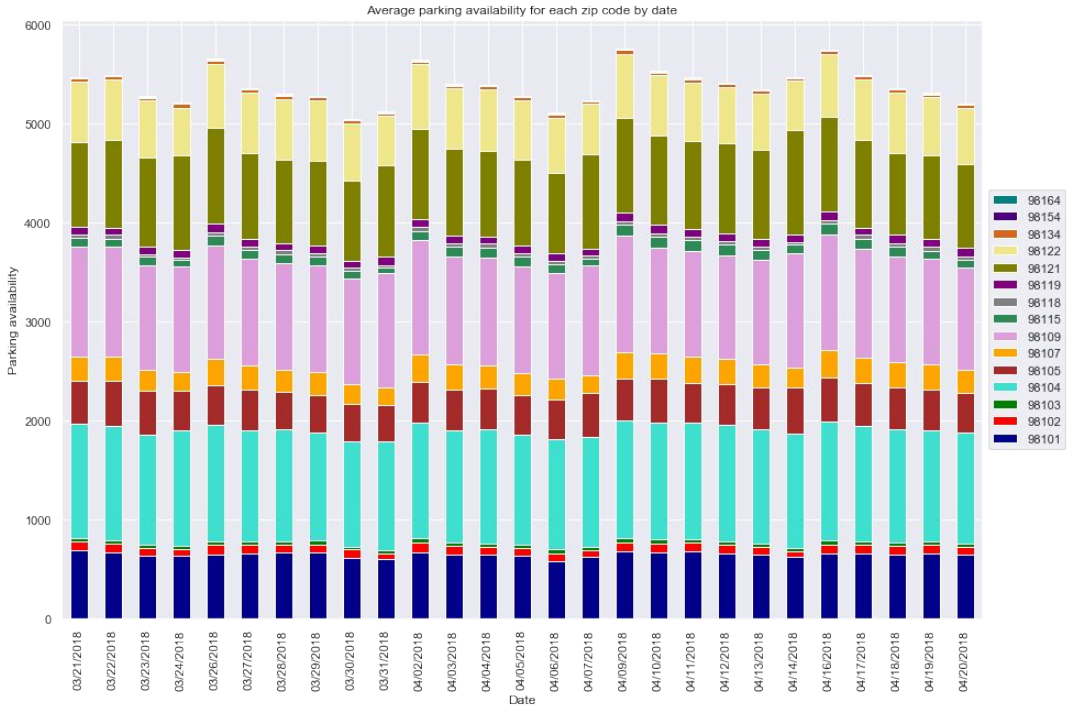
Looking at average parking availability based on location, in this case zip codes, we can notice a similar pattern emerge to the previous graphs. But in this case we are able to identify how availability varies throughout the day for each zip code and identify the areas where paid parking ends sooner.

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.



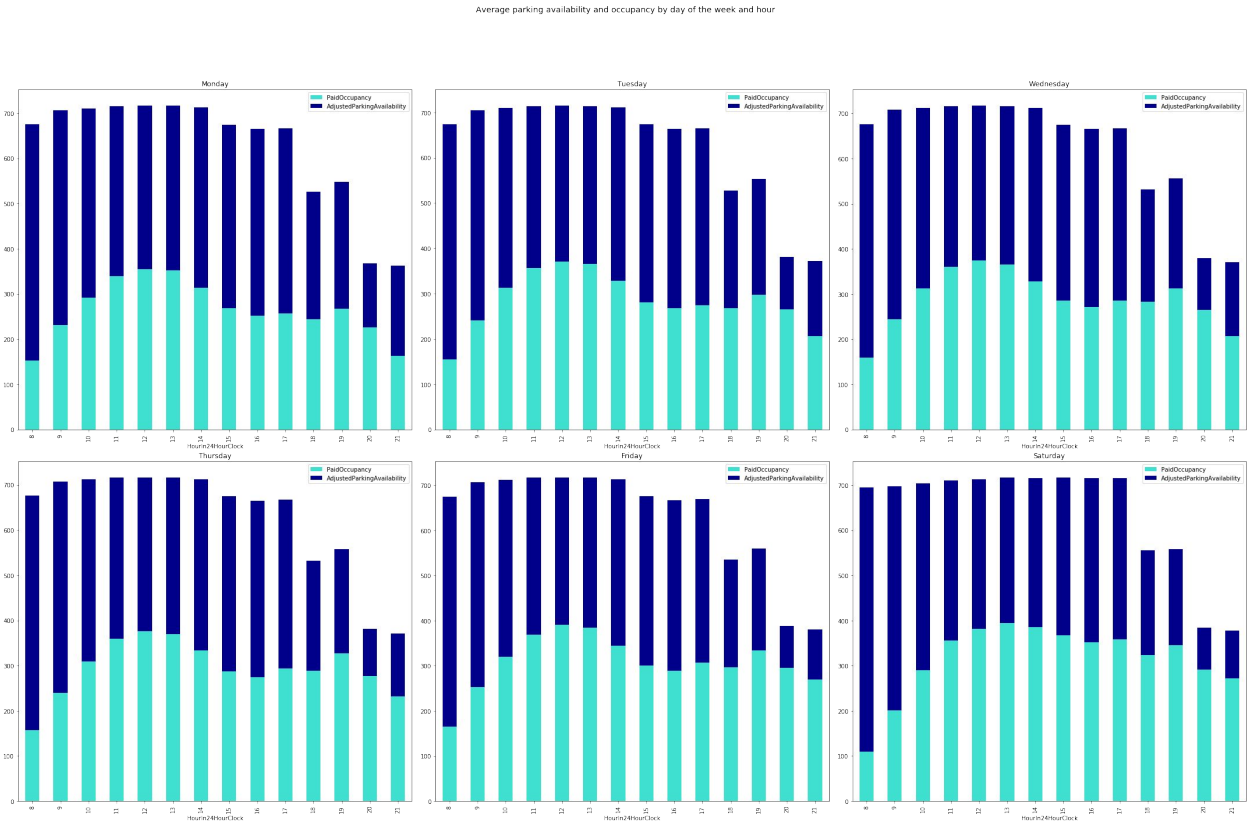
Patterns

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.



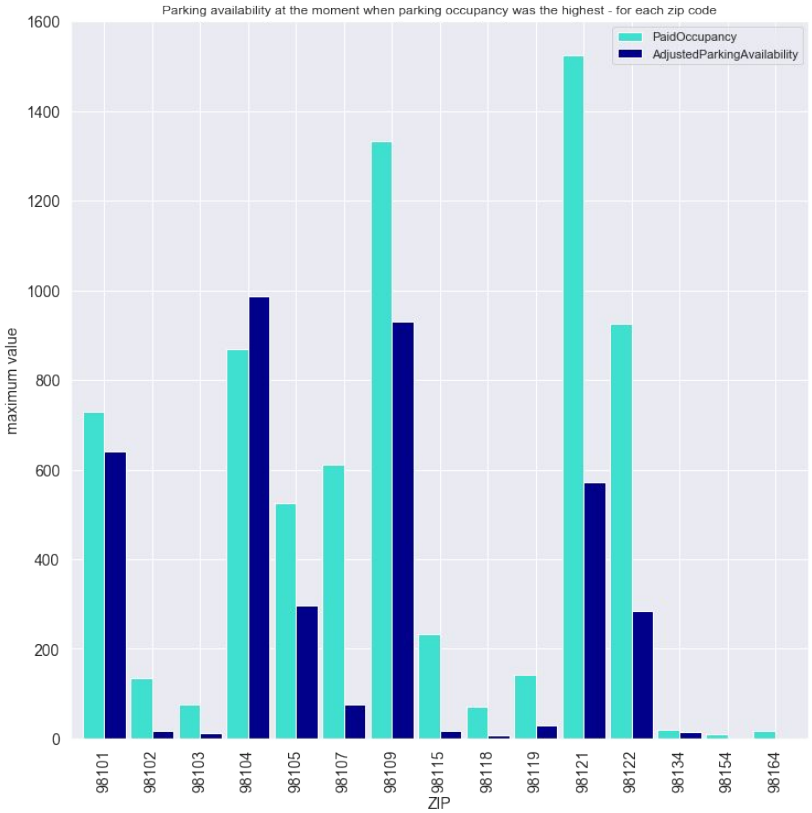
Patterns

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. Zip codes with less general availability were close to or at capacity at the time of their peak.



Conclusion and Further Exploration

We are able to establish that paid parking occupancy is fairly predictable for the city of Seattle and the explored time frame from 21.03.2018 to 20.04.2018. 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.

Conclusion and Further Exploration

For further exploration we would like to take a look at longer timeframe to identify patterns and trends across months and years. Furthermore we would have liked to explore the effects free parking has on paid parking, if that data were available. One additional aspect to test could be the effects of the price of parking on availability.

More details about this part of the project can be found here.

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).*

Techniques used

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.

More details about the techniques, methods and results can be found here.

Results from Linear Regression

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****, either* ***ZIP codes*** *or* ***Parking area****, but (surprisingly) not* ***Parking Occupancy****. Which is directly connected to Parking Availability.*

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

Results from 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.

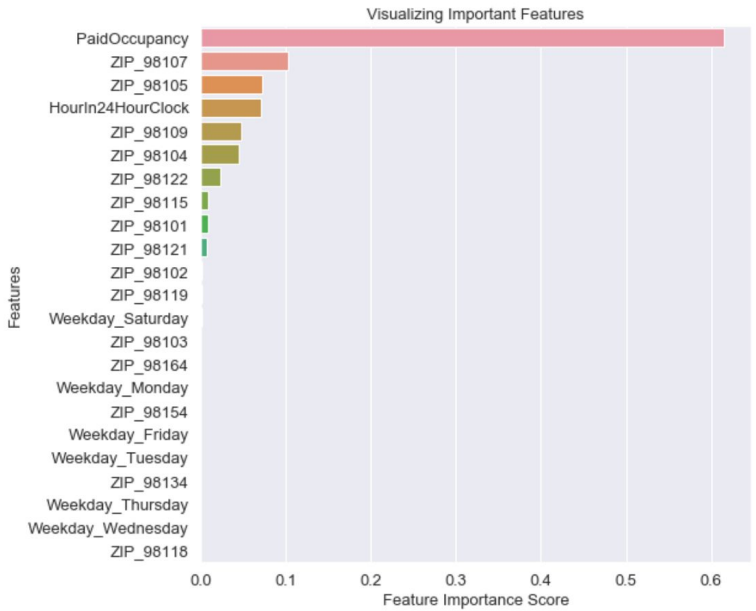
Going through all these models, we can see that the Random search model (marked green in the table), which was built with data based on parking areas with zip codes and parking occupancy, has the lowest errors. And the model ZIP code with parking occupancy has the largest errors and the lowest R_squared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

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

Results from Random Forest

*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.



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.

Conclusion

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.

Conclusion

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.