The Kronos Incident: Geospatial-Temporal Patterns of Life Analysis

Corey Vorsanger
COMP 4449

Repository can be found at: https://github.com/cvorsanger/COMP4449Final

Introduction

This is project is intended to showcase real world application of concepts I had learned during the Masters of Data Science program at the University of Denver. For this I choose one of the course projects; The Kronos Incident: Geospatial-Temporal Patterns of Life Analysis.

The analysis was a criminal investigation type deal. The problem posed was that in January 2014, the leaders of GASTech are celebrating their new-found fortune as a result of the initial public offering of their very successful company. In the midst of this celebration, several employees of GASTech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem.

GASTech assigned many of their employees company cars that could be used for both personal and business use. However, GAStech does not trust their employees. Without the employees' knowledge, GAStech has installed geospatial tracking software in the company vehicles. The vehicles were tracked periodically as long as they are moving. The vehicle tracking data for the two weeks prior to the kidnapping was made available for the analysis.

In addition to the vehicle data, law enforcement has been given access to the personal and business credit and debit card transactions for the local GASTech employees. Many of the GAStech employees also use loyalty cards to gain discounts or extra benefits at the businesses they patronize, this was also given to law enforcement. Once again, this data spanned the two weeks prior to the kidnapping.

The goal was to analyze this data to assist law enforcement by making recommendations for further investigation. To do these we needed to answer three questions. Firstly, what does a day in the life of a typical GAStech employee look like? Second, identify up to twelve unusual events and describe what makes them unusual. Lastly, because this data is imperfect, describe the considerations I took to address these imperfections.

Dataset

Being a course project that dataset can be obtained on the class GitHub at https://github.com/emmanueliarussi/DataScienceCapstone/tree/master/7_FinalProjects. The data contained is synthetic and depicts the daily lives of employees of GASTech over the two-week period.

The data is split into four different csv files; a gps file, car assignment file, a loyalty card file, and a credit card file. The gps file detailed geospatial temporal data based on car id. That is for each car it gave a longitude and latitude reading plus the timestamp that the reading was taken. The car assignment file tied these cars to the employee the car was assigned to. It gave the car id, employee name, employee job title, and employee job type. Lastly, the final two files contained purchase data for each employee during the two weeks. The variables contained where employee name, purchase price, purchase location, and timestamp of the purchase. The only difference was that the loyalty file contained only purchases the employees used loyalty cards for. The credit card file contained all purchases.

As you might have noticed much of the data is strings; employee name, job information, and purchase location. We did have some numeric data such as purchase price, latitude, and longitude. Finally, we had a datetime object in some of the files. There were some data points that were common across file, such as employee name. Table 1 details data types and in which file each variable was in.

Variable	Type	DataSet
CarID	integer	Gps, Cars
Employee First Name	string	Cars, Credit Card, Loyalty
Employee Last Name	string	Cars, Credit Card, Loyalty
Current Job Type	string	Cars
Current Job Title	string	Cars
Latitude	float	Gps
Longitude	float	Gps
Timestamp	datetime	Gps, Credit Card, Loyalty
location	string	Credit Card, Loyalty
price	float	Credit Card, Loyalty

Table 1: What is in Which File

Data Exploration

With the data read in it was time to get an idea of relationships, needed preprocessing steps, and possible feature engineering. This was the most important step in the data set and the one I spent the most time on. Without this step I had no idea of knowing steps I needed to clean that data that would facilitate my analysis. This step also gave me some ideas into areas I needed to look in.

I first set out to understand what was in the car assignments data. Particularly, I wanted to get a sense of the different employment areas at GASTech. I

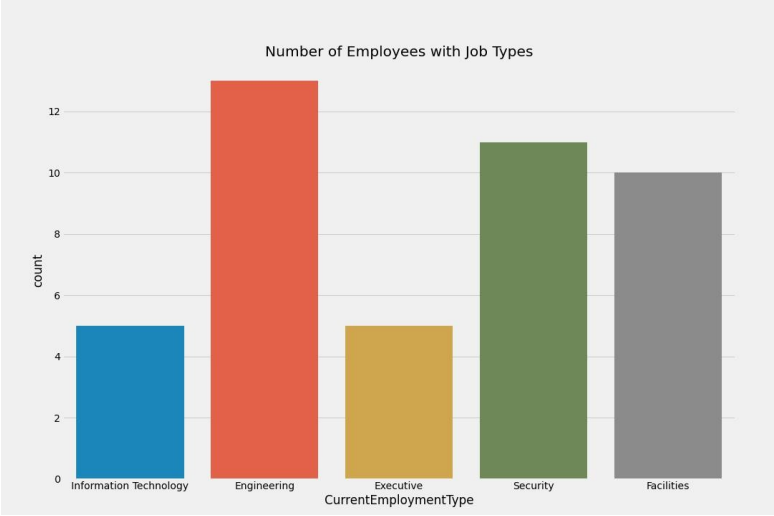

Figure 1: Amount of Job Titles per Type

found that there were five different employment types; IT, Engineering, Executive, Security, and Facilities. Engineering was the largest department with thirteen employees while IT and Executives had the fewest with five each. I also noticed that for the most part this dataset was whole with almost no null values. The only exceptions was that employees with the job title of Truck Driver had no car assigned to them. This was due to the fact that they didn't have a company car they could use day to day. They were assigned a truck that they were only allowed to use during workday.

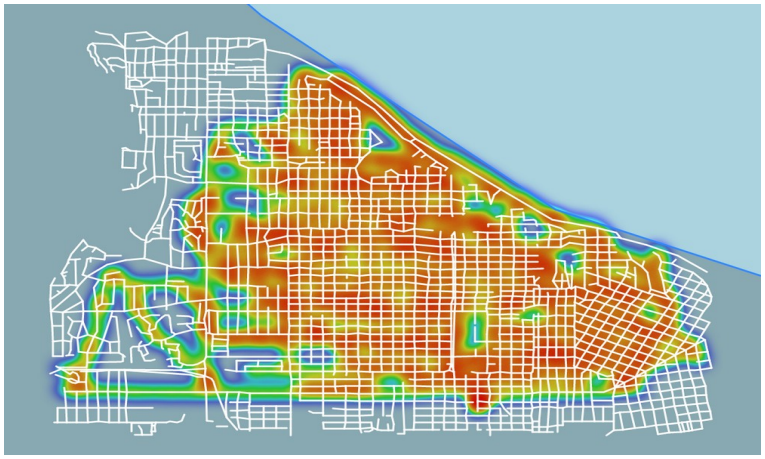

Figure 2: Heatmap of Where Employees Are

Following the car assignments data, I sought to find something out about the gps file. I easily found out that we had no null values. By using the median values of the longitudes and latitudes I could tell that most of the employees went to locations in the south east of Abila. This was confirmed by creating a heat map that plotted all of the gps data regardless of timestamp. You can really get an idea of where the employees hung out with this visualizations. You can see that people really liked the coast and the center of the city. This is where most restaurants, parks, and the golf course where located. You can also see where GASTech is located. It is the "bump" on the southern part of the heatmap. Lastly, you can see that while people went to the airport in the west it wasn't as frequent.

I then moved on to the loyalty and the credit card data. I explored these together because they contained much of the same information. I was surprised that file contained null values. The most valuable exploratory analysis function I did was to look at the distribution on prices. I found that in each most prices were $100 or less. It was also noteworthy that the distributions looked very similar, but the loyalty card purchases tended to be lower. One interesting point was a charge for $10,000 that was clearly abnormal. Finally, looking at the locations I found that restaurants and coffee shops tended to be the most frequented spots.

Figure 4: Loyalty Card Price Distribution

Figure 3: Credit Card Price Distribution

With the individual files explored I could start looking at the relationships between the files. For instance, I could look at differences in the location data by job type. This showed huge difference on how different employees spent their day. For instance, the heatmaps to the right show locations for people in the Facilities department and the IT department. The Facilities department went to the airport whereas the IT department did not. The IT department was also much more prone to being within the center of the city. There were also other differences that where highlighted when making a time dependent heatmaps. An example, is that employees in the Executive department were more likely to go home earlier than anybody else.

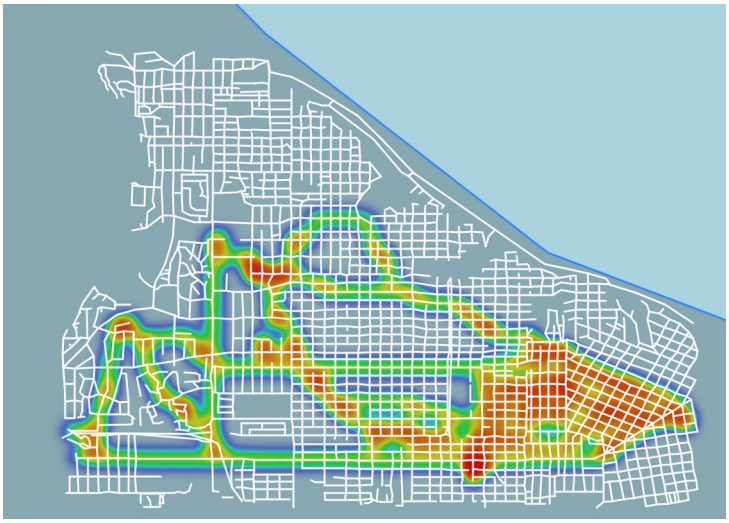

Figure 5: Facilities

Another difference that I found were activities on the weekend and weekdays. Weekdays tended to be work focused. During the middle of the day employees tended to be at work, at restaurants for lunch, or running work errands. During weekends however, the employees had more fun. While they still got coffee and went to restaurants most did not go into the office. Instead, they did fun activities like the parks, or play golf, or go to the museum.

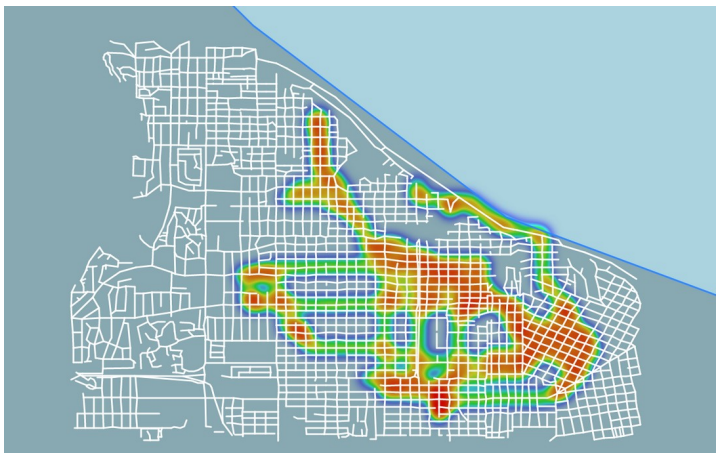

Figure 6: Information & Technology

Lastly, I wanted to look at the relationship between the spending patterns of somebody and the job type they had. I found that most morning and nightly spending patterns were consistent. Most employees went to get coffee in the morning and went out to eat at night. During the weekday Facilities was much more likely to be away from GASTech. Engineering also was also prone to being outside the office. I also found that Facilities had by far the highest spending of all job titles. I was surprised that the Executives didn't have the highest spending. I think this hints to the fact that Facilities had many company related charges, such as truck maintenance.

		price	
CurrentEmploymentType	**Is_Loyalty**	**median**	**mean**
Engineering	0.00	23.17	42.52
	1.00	19.62	34.49
Executive	0.00	29.22	56.61
	1.00	25.82	49.68
Facilities	0.00	1,135.20	1,647.12
	1.00	1,158.36	1,631.63
Information Technology	0.00	23.63	104.38
	1.00	18.31	29.12
Security	0.00	26.92	48.06
	1.00	19.80	38.82

Table 2: Spending per Job Type

Preprocessing

An important part to do before data analysis and machine learning is preprocessing to clean the dataset. This will allow us to manipulate data and to feed it into a machine learning models. The steps needed to clean this particular data is merging, filling null values, feature engineering, and encoding.

Because we received four different csv files, the first step is to merge the files into useable dataframes. While it is usually preferable to collect everything into a single dataframe, I decided to created two different dataframes. One was a location dataframe; made by merging the gps and car assignment files. This dataframe gave us longitude, latitude, employee name, and job information. This would allow us to analyze locations segmented by job titles or job types. Next I created a purchase dataset by combining the credit card, loyalty, and car assignment files. An interesting aspect of merging the credit card and loyalty files is that they contained a lot of the same purchases. We could identify the duplicate purchases by looking at employee, data, and the cents number of the purchase. To reconcile any price difference I took the loyalty card purchase. This was great because the loyalty card data was less prone to outliers. The final purchase dataframe detailed 1500 plus purchases. The relevant attributes detailed in it was purchase time, business, price, employee name, and employee job information.

The next step was to fill in null values. Both datasets had a few. For the purchases data there were some employees that didn't have a car assigned to them. Therefore, their job information was null. These entries still contained valuable information for our analysis so I decided to fill these null values with "Other". Similar the null values in the locations dataframe arose from the merge. The gps file had locations of five trucks that were not assigned to anyone. The trucks were driven by the hired truck drivers throughout the workdays. I then filled the null values for name and job information for these trucks with a name of Truck Driver with a job title of Truck Diver and job type of Facilities. Unfortunately, there wasn't an easy way to tell which Truck Driver was driver which truck at a particular time easily right away. Later in the analysis I tried to match purchase data with location data to infer who was driving the truck. I had moderate success at this but wasn't able to find any concrete recommendations through this.

Another sudo-null aspect of the locations data was the inconsistent nature of the gps readings. Sometimes they would be every second, sometime every couple of seconds, and sometimes nothing for hours. At other times there would be two data points with the same timestamp for the same person. To remedy these two issues I resampled the location data to one minute intervals. This reduced the possible data points for a person to one per minute. I was not concerned with a loss of data as people don't move too fast and a one minute resolution in the locations was good enough to track people to make conclusions about their daily lives. This still left some holes in the dataset when there wasn't a data point in a particular minute for somebody. I dropped these minutes and assumed that no change took place from the previous data point.

Next, it was onto feature engineering. I didn't find too much that I could make other than from the timestamp data for both the locations and purchases dataframes. For each I made a boolean feature that denoted if a day was a weekend. I found that there were differences in the employee behavior between weekends and weekdays, so it made sense to make this attribute available in the analysis. For better and easier querying and segmenting I separated the timestamp parts day, hour, and minute into their own features instead of being just being embedded into a datetime object.

Lastly, I had to deal with the categorical variables; name, job title, job type, and purchase location. For this I selected to use one hot encoding. One hot encoding was the correct choice here as there was no order to the categories, so it was better to make boolean features. For the purposes of reproducibility for machine learning I performed this step through a scikit-learn column transformer pipeline.

Machine Learning

After I had data that was nice and clean I could set out to try to find possible recommendations from the data. To accomplish this I first turned to machine learning. By using a unsupervised algorithm maybe I can easily identify some outliers to further investigate. Scikit-learn has two great algorithms that specialize in outlier detection; Isolation Forest and Local Outlier Factor.

Isolation Forest is a tree based algorithm that tries to isolate data points. The algorithm will first select a random feature to split on, and then it will select a random value to split that feature. This is repeated for a certain set number of time creating a Decision Tree like structure. The number of splits it takes to isolate an observations is considered its path length. This process is repeated creating a forest (much like Random Forest) and the path lengths of the observations are averaged. The observations with very short path lengths compared to the rest of the dataset are said to be an anomaly. Unfortunately, this algorithm didn't produce any results for either the purchase of the location data so the search went on.

The Local Outlier Factor algorithm is a k-nearest neighbor based anomaly detection model. Using knn, a density is calculated over the data's feature space. Then we can look at the densities at each point. The points that coincide with a very low density are then considered outliers. This algorithm worked much better! I was able to find six different location points and fourteen purchases that needed to be looked at. The model doesn't explicitly give you a reason why a point was flagged, and these points need to be investigated further to figure out why.

Analysis – Question 1

The first task was to describe common daily routines for GAStech employees. This would help us identify odd behavior by establishing what does a day in the life of a typical GAStech employee look like. I was able to make conclusions based on two may avenues. First and foremost, creating time dependent heatmaps. I created these for both the overall data and segmented by job type. With these visualizations it was easier to see common movements. I also looked at the most common purchases by hour. By looking at the type of locations I could also gauge the patterns of life.

The basic pattern I found was that coffee places were popular in the morning. From mid to morning tended to be in office or work related purchases by Facilities. At lunch-time we see many employees go to restaurant and then back to work purchases for the next three hours. At about 5:00 PM we start seeing employees going back to restaurants for dinner before heading home. This final part may be less accurate because of bulk transaction reporting done at the end of the day.

During the weekend we see people go to the golf course, parks and the museum to have fun. Much of the same restaurants and coffee places were visited on the weekends. One key difference is that nobody goes to the office on weekends instead choose to spend times at the above mentioned fun places.

Analysis – Question 2

Next it was time to find some anomalies in the dataset and make recommendations to law enforcement. I was able to identify twelve unusual events. I limited this to twelve recommendations but there were probably much more to b found with more time. In the below section I detail each unusual event describe who, what, when, where, and why it caught my eye.

The first odd behavior I found was Isia Vann on the morning of the seventh. Looking at places Isia Vann went to everyday I noticed that they went to Brew've Been Served on every weekday but for one, the seventh. I also noticed that usually he went with Edvard Vann. Edvard was able to go on the seventh but not Isia. Looking at the locations dataframe it did seem like Isia was out and about and not just sick at home that day. This deviation from normal daily routing is highly suspicious and I am confident this is something that authorities should look further into.

Next I noticed something odd with the company president, Sten Sanjorge Jr. On the eighteenth Sten went to the Chostus Hotel and had a very high purchases. Other purchases here were around $100, but his was over $600! Additionally, this was the only weekend purchase at this location. Other employees, Isande Borrasca and Brand Tempstad, had multiple charges here, but this was Sten's only charge. This could just be the President of the company treating some of his employees and friends to a good time, but it does stand out. With it having three different elements of interest I have a high confidence that this should be further looked into.

Claudio Hawelon went to the airport on the tenth. This was odd because whil other truck drivers went to the airport multiple times, Claudio only went this one time. Additionally, he had the only Friday trip across the two weeks. He also was one of only two purchases that did to use a loyalty card as well. The time he went was odd as well, as it was a few hours before the other truck drivers typically go. I am highly confident that this should be investigated further. While he could just be filling in and helping somebody out there are too many suspicious factors in this trip not to ignore.

Lucas Alcazar had two separate odd behaviors that I identified. Firstly, on the thirteenth had a $10,000 purchase at the auto shop. This was much higher then any one else. Also, the other high purchases tended to be employees in the Facilities department. Lucas though was part of IT. I am 50-50 on this suspicion. While it is very odd and high it simply could have been a bad entry on the purchase amount and maybe it could have been $100. However, there is nothing to indicate that this is truly a false data point. Also, on the thirteenth he made an odd purchase at Daily Delaz. This was the only purchase at Daily Dealz though it was only for a couple of dollars. While this is defiantly odd, I am not overly confident that something will come out of investigating this further. Daily Dealz could just be a place to get cheap everyday items. The fact that the purchase was at six in the morning is not too worrisome as he did have an earlier purchase as well. This behavior should be kept in mind, but more promising leads should be looked at first.

On the eleventh Ingrid Barranco mad multiple visits to Hippokampos. This was odd because usually Ingrid went to Hippokampos once a day. However, on this day she went three times. What was the reason for going three times to this place in one day. I am moderately confident that this should be investigated further. If she had gone only twice there would be more doubt. Maybe she just had a massive hunger for this place that day, maybe there is a duplicate transaction that I didn't detect earlier. But three visits, is very suspicious. Was she planning with somebody that works here?

Hideki Cocinaro made mutiple purchases at Kalami Kafenion on the eleventh. He doesn't frequent Kalami Kafenion however on the eleventh he visited there twice. For one of these purchases that day he did not use the loyalty card. Of the four total time he visited this was the only time he didn't use a loyalty card. What's more, the purchase was higher than normal, and at an odd time; six hours later than is other purchases. I am fairly confident of this observation. The high price could be just due to him not using a loyalty card but the fact that he visited Kalami Kafenion twice on the 11th including once that was at an abnormal time is suspicious.

One of the few purchases at Frank's Fuel was Felix Balas on the eighteenth. This was one of two purchases there. Frank's Fuel is located at an odd place on the island for Felix. Firstly it is in the west where not many engineers go to. The purchase was also on the weekend, and it is not located near any restaurant, coffee shop, or leisure place. He could just really needed gas for his car and that was the closest spot to his house; most likely in the north but further toward the coast. Still very odd behavior, and I am moderately confident this could be something more and should be looked into further.

Nils Calixto decided to go to Katerina's Cafe twice on the seventh. It did do this one other time, on the thirteenth, so maybe he just really loves this spot. The most curious thing about this purchases I am highlighting though is that this purchase was at 1:30 PM. All of his other purchases here were between 7 PM and 8:30 PM. This huge discrepancy in time is a little concerning. However, I am not overly optimistic on this one. This is a place he likes to go to, and maybe he just wanted to try their lunch menu instead of dinner.

On the tenth Varro Awelon missed his daily stop at the Coffee Shack. I am assuming this happened before work, but all purchases were uploaded in bulk at twelve hour periods. I checked to see if he was anywhere else during the morning of the tenth but I didn't find any purchases and the location data didn't seem overly odd. While the fact that he deviated form normal behavior is odd, he could have just been running late that day however. Other identified leads should be investigated first.

Lars Azada had an odd behavior when he missed his daily coffee stop before work on the fifteenth. While this could indicate that he was off scheming at that time, I find the likelihood to be small. He could have just been running late that day. Other identified leads should be investigated first.

Finally, on the tenth Willem Vasco-Pais made a purchase at Hippokampos at an odd time. Additionally, he went twice that day! Typically he would go after work for dinner. However, on this dat he also went at lunchtime as well has his usual dinner stop. My confidence on this observation resulting into something concrete to the investigation is low. He could have wanted to try the lunch menu instead of always just going during dinner.

Analysis – Question 3

This dataset proved to be especially dirty. There needed to be thought on how to merge the data, handle inconsistencies in both purchase times and location times, and how to handle outliers. Without careful consideration of these factors our analysis could not be trusted.

For the merged locations dataframe. I had to consider a few things. One such thing was the varying time deltas between data points. I was able to find this by inspecting the time-lapse heatmaps. By looking at the spread and color intensity it was easy to see that there are a varying amount of data

points per hour. Digging deeper and looking at a single individual confirms that sometimes the location data was reported every second, sometimes once a minute, and sometimes there is nothing for hours. Reconciliation of this aspect was done through resampling the data. For each person i resampled their location to one minute intervals (the file size was too big for GitHub if I did a one second sample). This allowed the time deltas to be more consistent at one minute. However, there were still "holes" sporadically during the day and at night. An easy solution could be to backfill null values with the previous location. However, this would greatly grow the dataset without adding any valuable information and potentially skewing the data. I therefore decided to do nothing about these holes.

The resampling also helped for the situation where two location records for a person at a single timestamp. Resampling would automatically combining these into one data point.

Another imperfection I came across was the fact that the truck drivers had to share vehicles. This was due to the fact that they didn't have a company car they were allowed to use for personal use. However, we were given information about the five different trucks that were shared. We could assume that these were being driven by someone with a Truck Driver job title. If someone other then a Truck Driver were to drive one of these vehicles, it would be suspicious and would show up graphical on one of our heatmaps. It was possible to investigate purchase data in order to ascertain who was the driver.

Lastly for the location data I had to consider what to do with the outliers. Using a violin plot I see that there were some gps points considered outliers. However, they were not so extreme to hint that there was something off with the readings. Because the point of the project to identify outliers in order to make recommendations, the outliers contained valuable information that shouldn't necessarily be altered or dropped. For this reason, I decided to keep any potential outliers that way I could inspect them and determine if they were real and relevant to the investigation.

The purchases data wasn't without its headaches as well. As I previously mentioned I found during EDA that there were many of the same purchases detailed in the loyalty card and the credit card files (about 70% of the transactions). I was able to detect this by sampling the data for a single individual and comparing the two sties of purchases. I did find two differences between the duplicate purchases. I also noticed that sometimes the dollar amount of the purchases were different with the credit card prices seemingly much more prone to outliers. Additionally, the loyalty data just gave you the day while the credit card dataset also gave you the time of the purchase in certain situations.

To solve this I did a custom merge. I created a finction that went through each transaction and identified the duplicate purchases by looking at name, date, and cents amount of the price. I then kept the credit card timestamp as it was more detailed and the loyalty card price; more on why in the next bullet point). I kept track of all of the loyalty purchases through a boolean variable that was true if the purchase was detailed in the loyalty file.

As mentioned I took the loyalty card purchase price when combining the two files. This was due to outliers in the pricing. I was able to identify these through violin plots and in the inspection mentioned in the previous section. An interesting fact and the reason why I choose to hold onto the loyalty price, was that the credit card dataset tended to be much more prone to these outliers. The outliers were trickier to deal with as the whole purpose of this analysis was outlier detection. Deciphering between actually outliers (suspects) and bad data was nearly impossible. I therefore, elected to keep them as they might contain valuable information relevant to the investigation.

Another interesting observation I found was the time that purchase transactions were dated. As mentioned, the loyalty card file only gave the date of the purchase with no time. The credit card file also gave time however, it seemed that only some transactions were reported in real time. In other instances the transactions were updated in bulk, typically at the end of the day. Additionally, looking at a sample of some the purchases at an individual place, sometimes they reported transactions in bulk sometimes they were reported in real time, and there didn't seem to be a pattern. I had to keep these two factors in mind when analyzing the purchase data. There wasn't much I could do about them, so when looking at the time I had to have some doubt about the time.

Lastly, there were more employees with purchases then there were assigned a car. I found this by looking for null values after combing the loyalty car, credit card, and car assignment files. These employees were among a handful that didn't car assigned to them and didn't work as a truck diver. Unfortunately, I wasn't able to get the job information for these employees since this information was all in the car assignments file. These were still valuable data points so I couldn't drop them. To handle this I filled these null values with the value "Other". This allowed me to still analyze these points and to encode the job information before utilizing machine learning.

Conclusion

In the end we were able to define the common behavior of the employees at GASTech, identify twelve odd behaviors worth further investigation, and reconcile the imperfections in the data. We found that the employees loved their coffee and eating out for lunch and dinner. Then they did some work before going to lunch then more work followed by going out to dinner and going home for the night. Weekends were a bit more fun with employees going to the golf course or museum instead of going to work. The three most interesting behaviors I found were; the President having a very high charge at an odd time at the Chotus Hotel, Isia Vann missing her daily pre-work coffee stop with her husband, and finally Claudio Hawelon visiting the airport on a day truck drivers usually don't go there.

This was a great and rewarding project. I was able to learn techniques on how to manipulate and analyze geospatial-temporal data. I found the visualizations can be extremely powerful when digesting data. I was also able to get some practice at outlier detection. I also learned some of the strengths and weakness of using machine learning, visualizing, and simple pandas functionality as they pertain to data analysis. I was very surprised of the power of visualizations while looking to find answers to the proposed questions.

While I was able to identify twelve odd behaviors, my list was not the final product and there is definitely room for improvement in this project. The biggest factor for possible improvement is of course time. More time would allow for more investigation into machine learning algorithms that for geospatial data than the ones I tried. I would also try to look at the relationship between purchases and locations. Where there any purchases and location mismatches? This could also help us find odd behavior.