First Report

Zeyu Chang, Kan Zhou, Irene Chang

Slide Link:
https://docs.google.com/presentation/d/1_3b_YXcoQimE5bzJdXNNyKWnhJ5Js52uCLwsSI8Wsao/edit?usp=sharing

Abstract:

The topic of this project is the first mini challenge in the VAST Challenge 2017. We decided to go with this project because we think the analysis on human impact on wild animal's population has lots of inference potential since this has been the topic for a long time and we think it would be a great idea if we can tell this age-old story from a data point of view. Therefore, we aim to build a system where all elements are connected to one another.

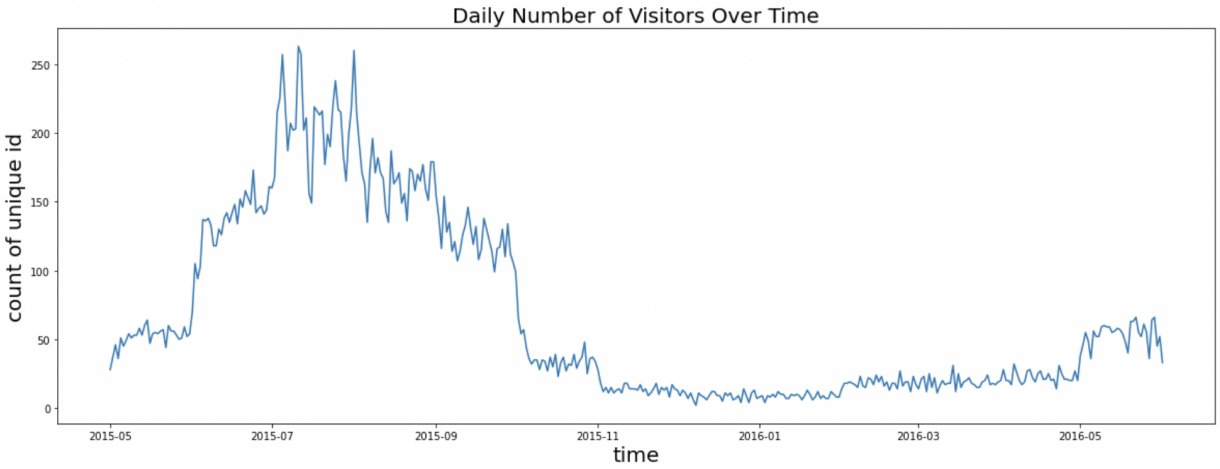
The goal of the project is to build a system that helps identify odd behaviors of the vehicles visiting Boonsong Lekagul Nature Preserve, from which we better understand what causes the decrease in bird population in the preserve. Our system focuses on identifying patterns of the passenger flow within the preserve over time that stand out among all the visitors. We also aim to provide visualizations that showcase information on both the broad and the more detailed level, the latter of which is drilling down on the odd observations to further investigate what may be the cause behind it and to generalize the behaviors of such vehicles in order for the people in charge to pay more attention to such odd behaviors in the future. The data provided for us has information on the type of vehicles and types of gates in the preserve, as well as the time recorded when a vehicle passes a gate, all of which we will use to analyze patterns among the day-today traffic in the preserve. More interestingly, there is a gridded map of the preserve in a bmp file, which we plan to find a way to build a part of our system out of.

Description of Data:

1/ By conducting an exploratory data analysis for the given dataset, we have the overview of the data:

	Timestamp	car-id	car-type	gate-name
count	171477	171477	171477	171477
unique	123133	18708	7	40
top	7/30/15 11:54	20154519024544-322	1	general-gate7
freq	9	281	67698	16119

2/ Then, we plot the number of visitors each day over the given time period:



We can see that visitors from May 2015 to Oct. 2015 are significantly high and that could be a reason why the bird population decreased.

Goal and tasks:

1/ Build a system that visualizes information on a general level.

For this goal, we want to utilize the bitmap image provided to visualize the dynamic of the traffic over time, as well as to highlight the vehicles that visit the routes which they are not supposed to go in. The general view on the passenger flow not only helps us identify which vehicles violate the rules, but also gives us a sense of the specific period in a year, for example, that more occurrences are observed, which we think is more intuitive than time series graphs with multiple lines.

In order to achieve this task, we have to research a way to process the bmp file based on the colors representing different objects in this graph. Ideally, we want to use gradients to show the traffic density on specific routes at a given time.

2/ Visualize first to identify interesting features/observations

There are not a lot of possible combinations of features in the dataset but it is essential that we extract only those that show important information and can be connected to one another.

In order to achieve this task, we are experimenting visualizing different combinations of features. So far, we have been able to identify that there are several patterns that are worth taking a closer look into. Yet, more exploration is needed to fully understand what way we can use the relationships between these variables to make useful visual systems.

3/ Feature engineering

There are variables such as time data which we think would give us a lot of insights if we can extract different levels (time in day/month/season/year) or calculate the length of stay, for instance, and visualize the trend in traffic, number of people visiting or breaking the rules over time. This step is important in organizing the components of visualization in our system as well so it needs to be done early.

4/ Use of filtering to combine several graphs into one

In order to maximize the use of space on the system, as with previous visual analytics systems shown in class, we need to figure out a way to use interactive filtering to display multiple insights within a graph. The big challenge of this task is figuring out what components can be combined in a graph, meaning they should help achieve a common goal and show a common point of information.

To achieve this, we have to explore all the major features/relationships in the data first. Then, we need to generalize the parameters in the visualization. Selecting the appropriate tool is another challenge. We are leaning towards employing vegalite to display our system, but Tableau and RShiny can be the alternatives if we have any problem with vegalite.

5/ Other numbers and statistics, tell a story

It is important that we keep in mind the connectivity factor in designing the system, since a thoroughly connected visual analytic system will make our model much more interpretable. This includes making use of filtering, colors, tooltips, and, on top of all, deciding the appropriate tool to use.

In finalizing the report, we also need to include the statistics and the numbers on the interesting insights we found while doing the exploratory analysis since not all information can be included on the system, while the report should be prescriptive as well.

Hypotheses

According to the discussion above, we conducted the following Hypotheses:

- Birds are decreasing in population because there were more visits in 2015 summer/fall.
- Cars that violate the rules tend to do so at specific points of time in a day and it is abnormal.
- The pattern of length of stay(too short/too long) and visit frequency could be unusual for some sites.

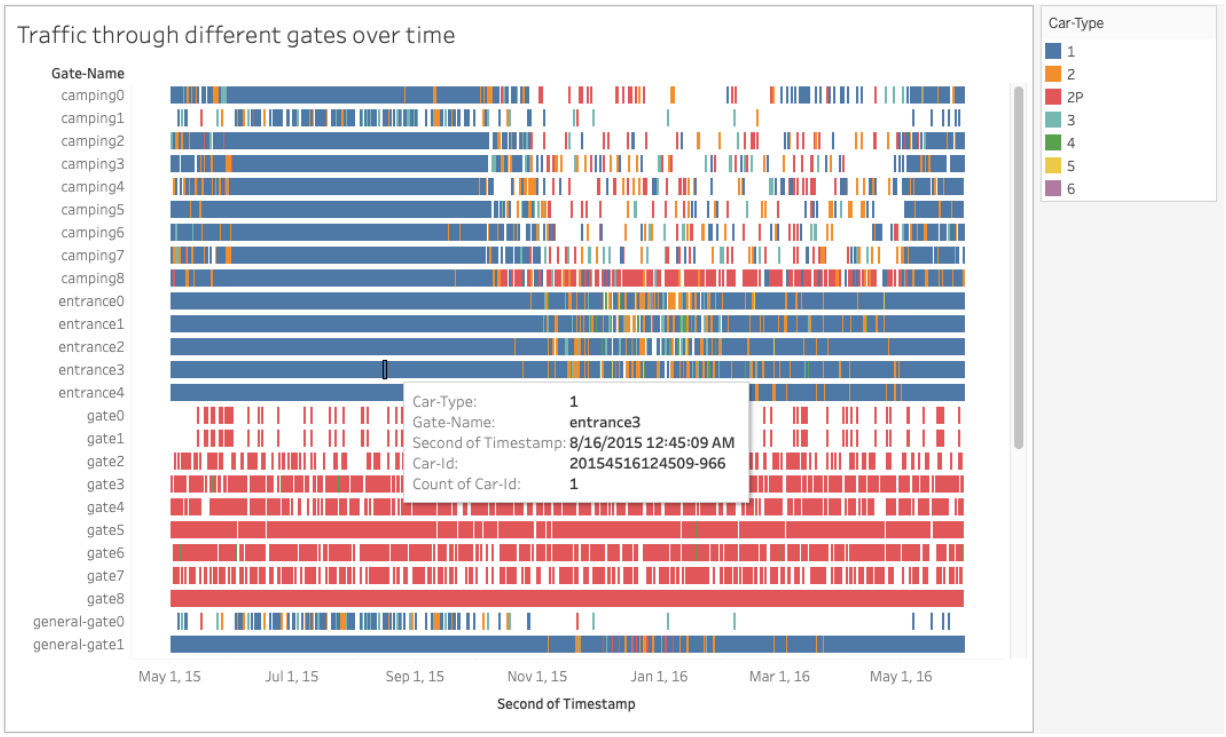
Exploratory analysis, storyboards for hypotheses

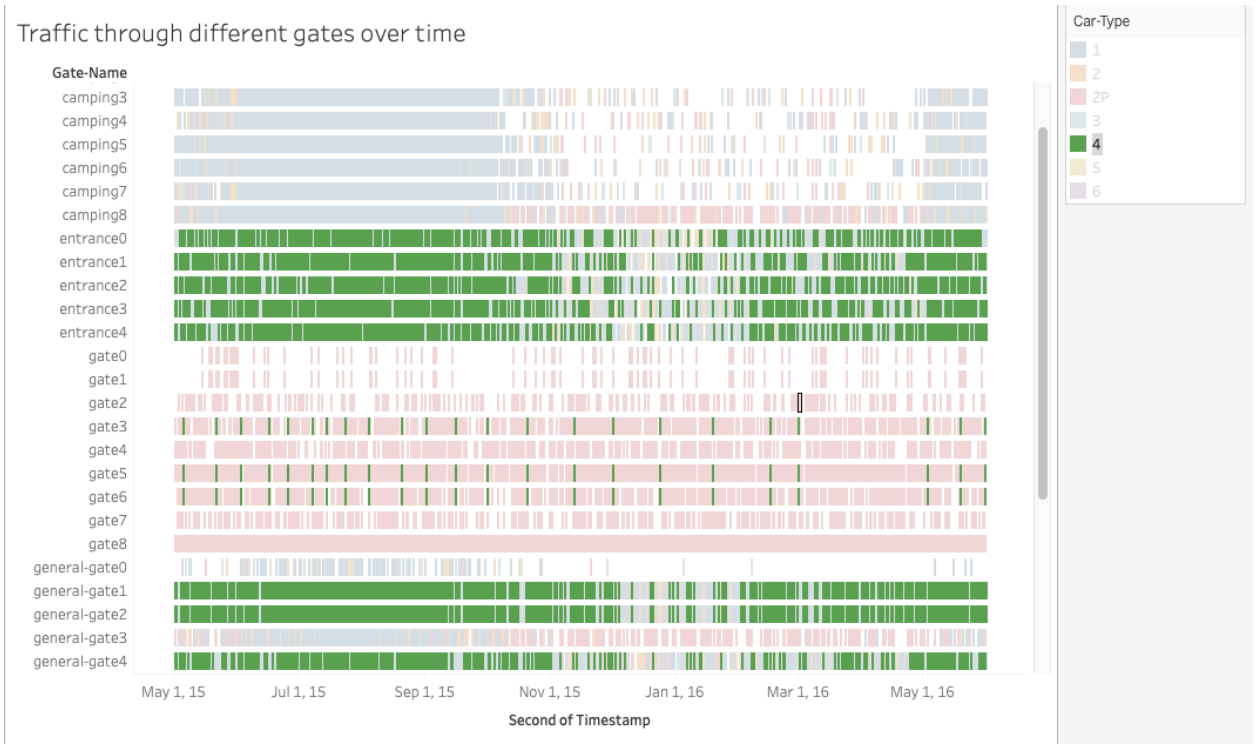
To develop a storyboard and decide on the design of the system as well as on which features to include, we use Python and Tableau as the main exploratory tools. The process helps us identify interesting insights that answer our initial hypotheses.

1/ Feature engineering:

- Gate type: Instead of having each gate being its own category, we extract the information of what type of gate each sensor is (entrance, camping, general-gate, gate-stop, ranger-stop, or ranger-base). This helps in generalizing the pattern across the main types of gates, rather than at each individual gate.
- Duration of stay: We picked the last time sensor detection and first time sensor detected the car id and calculated the time difference for all

2/ Formulate the behavior of different vehicles:





To further explore this dataset, we visualize the relationship between the number of vehicles over time by gates. We aim to maintain the features from Tableau when transferring the graphs

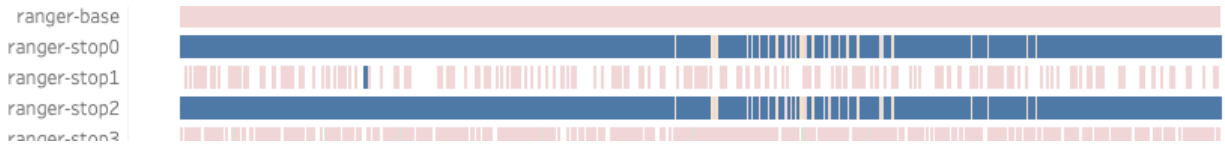
to vegalite, which will include the filters for car types, car ID, and the interactive tooltip that include information about the vehicles.

We found that, as expected, the rate of visits during summer is much more dense than in the fall. The Gantt view with filters on the types of vehicles not only allows us to see the density of visitors, but also to have a sense of how different types of vehicles travel within the preserve on a day-to-day basis. In other words, it gives us a sense of what can be considered conventional in the traffic flow:

- The 2P vehicles seem to be the Preserve Ranger and only they can cross the “Gates”, as well as ranger-stop 1, ranger-stop 3-7, none of them is registered to have passed the Entrances.
- The vehicles of type 1, 2, 3 are campers, they enter the camping sites most frequently, and share similar patterns.
- The type 4, 5, 6 vehicles don’t go to the camping sites and according to the map image provided, therefore they don’t pass general gate 6 and 3 (both of which lead to the campsites), and share similar patterns which are quite different from vehicles type 1, 2, 3. We think it's because the big vehicles have to park elsewhere and are not allowed to enter the camping places.
- For 2P vehicles, they pass camping8 and camping5 the most, but only in camping 8 do we observe the scattering pattern of visits over time, in contrast to camping 5, which appear to concentrate in just some period of time (maybe they usually go in groups over a short amount of time).

3/ Identify those odd-one-out:

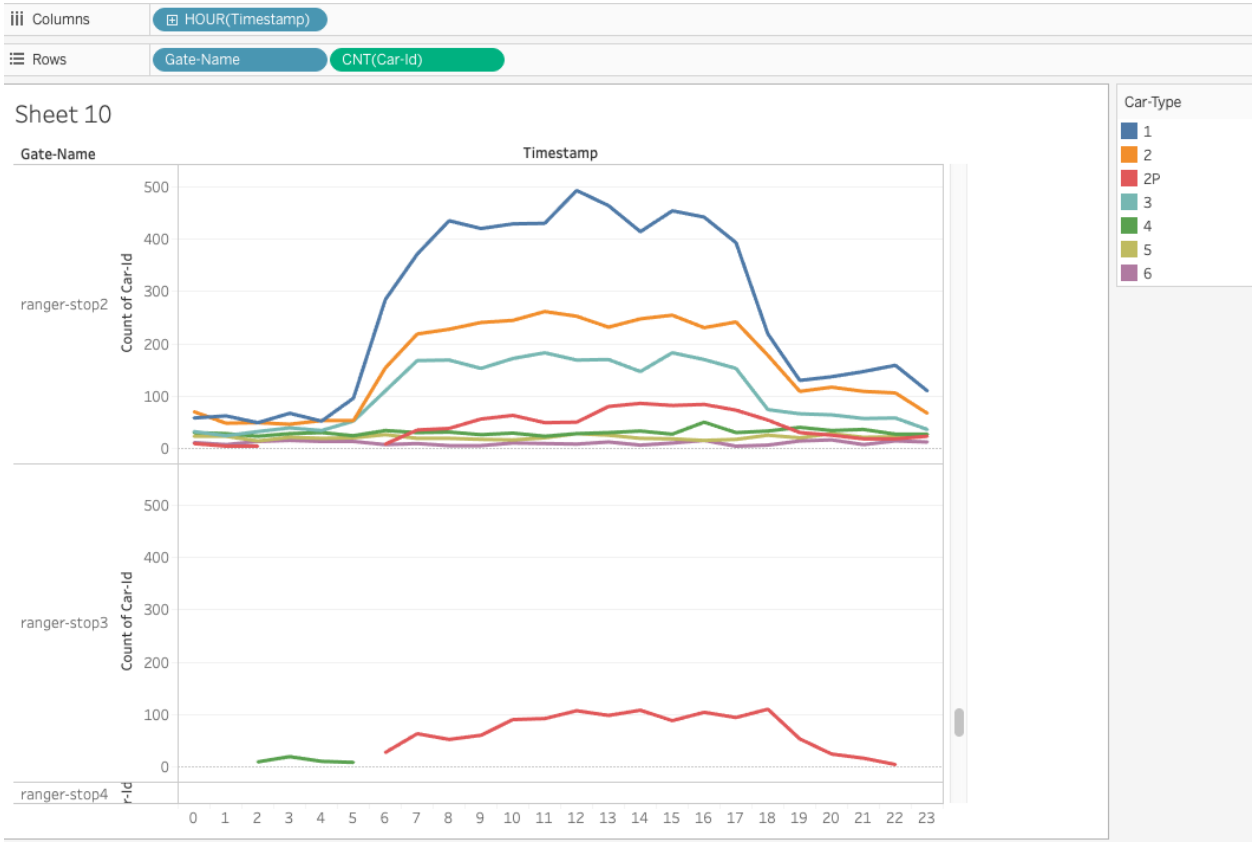
Not only does the Gantt view let us observe the general pattern, but it also allows us to identify strange behaviors that stand out:



- There are several instances of vehicles type 1 entering ranger-stop 1, which they apparently are not allowed to.

- At ranger 3, 6 and gates 3, 5, 6 there appear to be vehicles type 4 passing the sensors, which they are not supposed to.

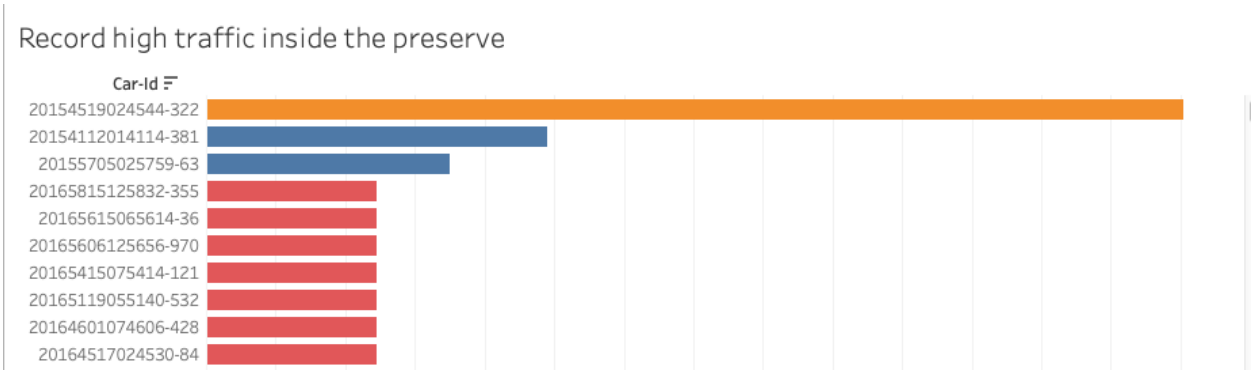
After finishing with the general observation, we start to dive more into more details and start exploring individual patterns and behaviors, especially those with strange travel patterns.

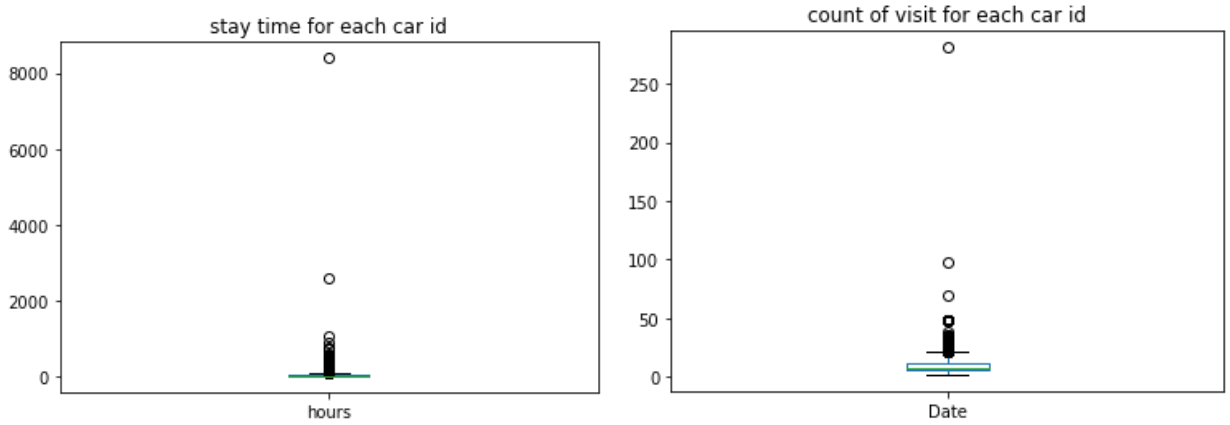


- By the second point in our hypotheses, we employ line charts to see the activities of different vehicles at different hours in a day, a snapshot of our finding is shown above. The normal behaviors are consistent within the gate type and vehicle types. However, in the ranger-stops, as expected from previous findings, there are instances of vehicles type 4 entering this area. What is more notable is that these vehicles enter these areas at very odd hours (2am-5am)

4/ More vehicle-ID-level analysis

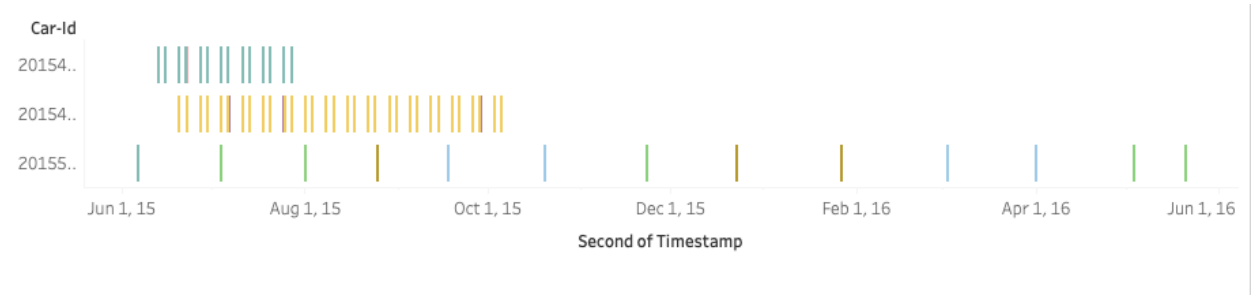
Firstly, we examine the vehicles that go into the park a lot

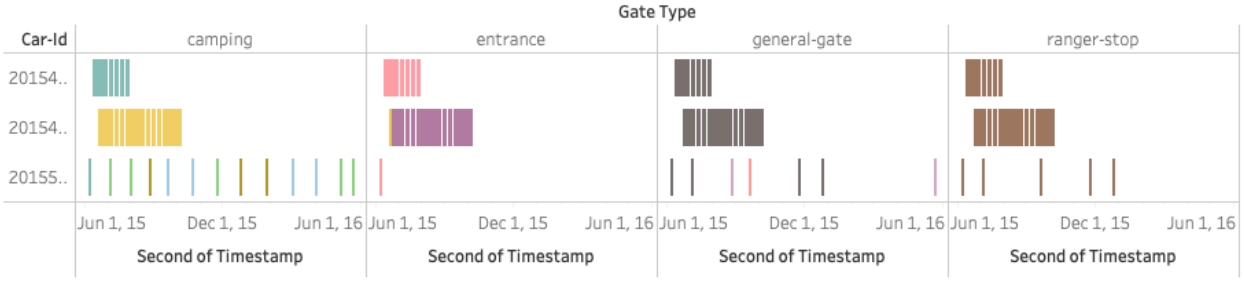




- We next explore the third point in our hypotheses, and we find out some anomalies in the length of stay the number of visits. Most of the cars that travel a lot inside the preserve are mostly Preserve Rangers, but there are 3 vehicles that had record high travel inside the park (2 type 1, 1 type 2), the vehicle of type 2 visited the park for around 280 times.

- In addition, some vehicles just enter and leave the preserve without visiting any other sites (only passing the entrances), and some just stay inside the preserve within 1hr





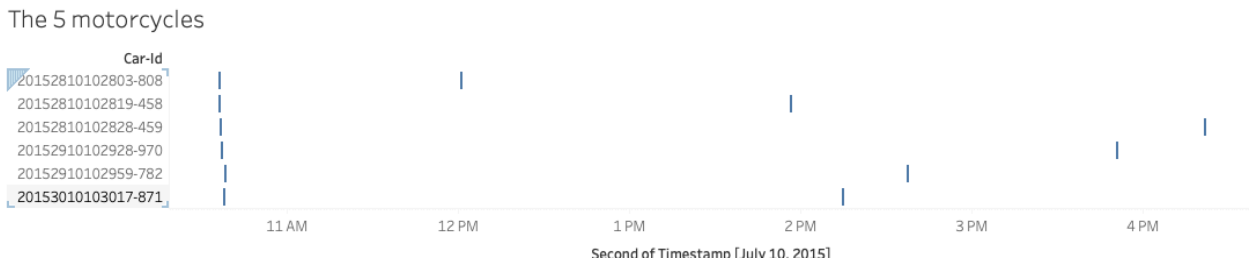
A closer look using Gantt view and color-coded based on the gate name suggests that:

- The 2 vehicles of type 1, 20155705025759-63 and 20154112014114-381, visit the preserve for a lot of time, and their length of stay are the same every time, and this pattern continues for a few months.9
- The car of type 2 (322) visit the park every 4 days, stay for 3 days, from june to october
- For 322, they always pass these sensors (camping4, entrance4, except for entrance 1 once, different general-gates, ranger-stops 0, 2). On the other hand, for 381, they go through entrance 0, to camping 6, general-gates 5,2,1, and ranger-stops 2, 0.
- 63 seems to go to a wide range of destinations over the period of a few months. It seems to suggest that 63 never leave the preserve, and it just relocates to another destination once or twice every (few) months.

Next, we examine the vehicles type 4 trespassing:

Looking into the table of these type 4 vehicles in the ranger-stops, arranged by time, we saw that they always go the same route (gate 6->ranger-stop6->gate5->gate3 -> ranger-stop3 ->and back in the reverse order ->out, for all of these type 4 vehicles)

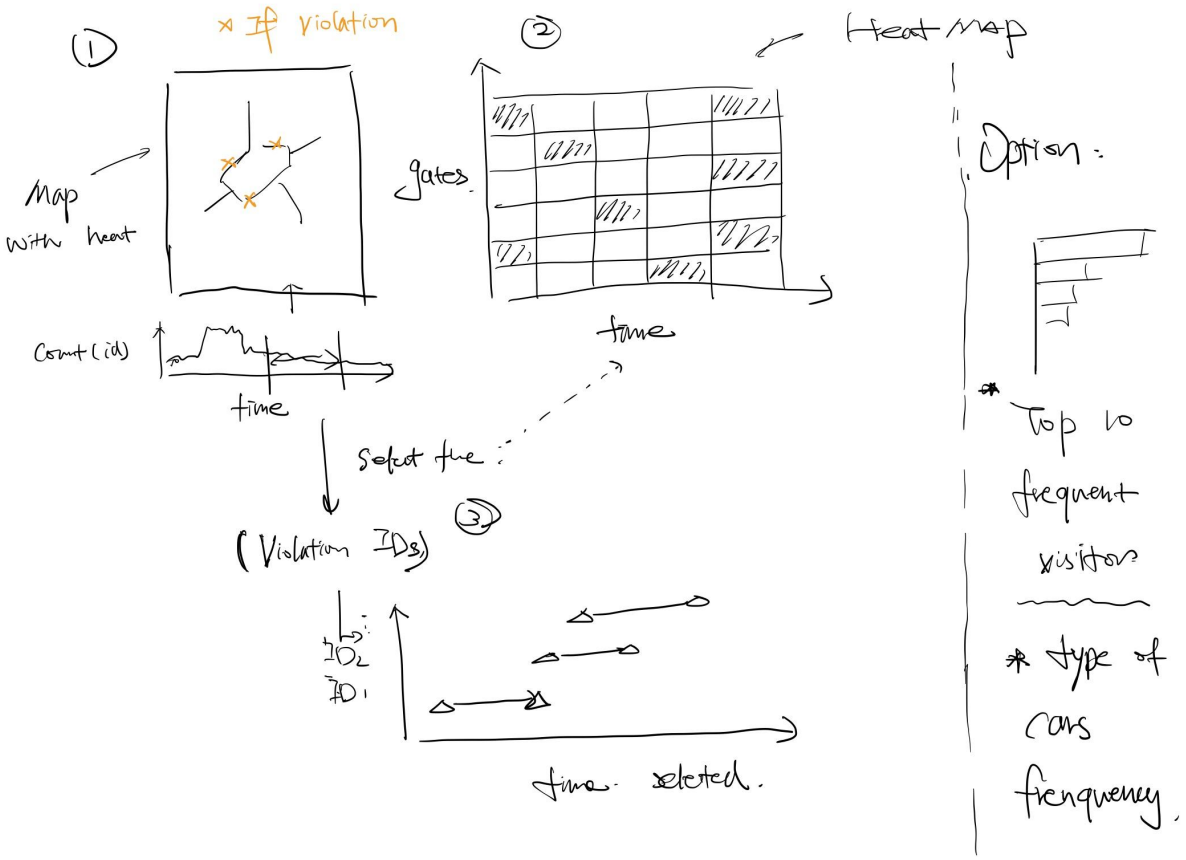
Finally, we examine the type 1 vehicles entering the gate:



On July 10th, there are 5 motorcycles going into the ranger stop 1 area which they are not supposed to at approx the same time. Maybe there is a drive for going into a restricted area in a group.

Sketch of the system

Version 1.0: Initial design

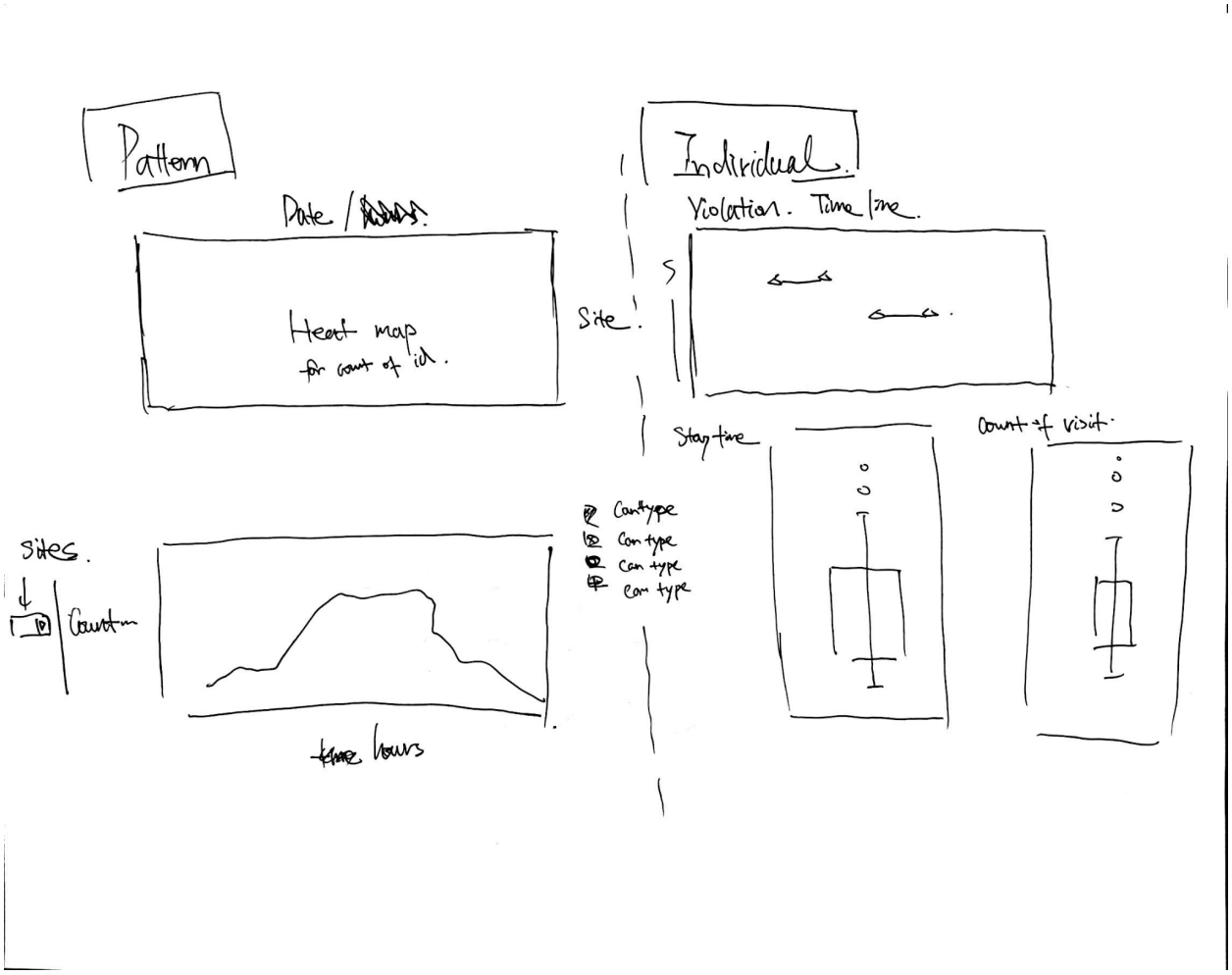


Version 2.0: Updated design, final layout

In this version, we changed the structure of the layout after carrying out EDA on the dataset in order to better reflect the findings on a wide range of visualizations. The layout is now finalized

to have 3 graphs whose general purpose is to give an overall view of the activities in the preserve, from which aid our team in identifying the strange behaviors of the visitors of the park, who are the possible culprits for the decline in bird population.

Besides, there is a visualization whose purpose is to drill down on the individual level of the vehicles that we think are the potential culprits and observe their travels within the preserve. This feature is represented by the timeline structure which can be zoomed in and out on scrolling, which helps us see the route of each suspicious behavior. Ultimately, this helps us understand more the motivation behind each odd behavior and find a way to handle it.



Project timeline

Today is March 25, and we include our goals and steps in our report. We described some problems statements and hypotheses, based on the data visualizations we draw, there are

some simple conclusions after the hypothesis. Before April15, we will have the detail solution for questions like specific data graphs, descriptions and the explanations of our sketches. How we make the analysis step by step. Also the conclusion of our problem statements like building a system to identify odd behaviors of the vehicles visiting Boonsong Lekagul Nature Preserve, from understanding what causes the decrease in bird population in the preserve.

From phase 2 - phase 3:
Split the work to implement four graphs with their own interactions individually in two weeks. Due by April 30.
After conducting the graphs, we are going to add interactions among those diagrams to increase the usability of the system.
By May 11, we will have a well prepared project report. A abnormal-detecting visualization system and a presentation slide.

Progress Update

From March 25 to April 15, we came up with a detailed final draft for the design of the visualization system. We also specified the interaction element in four main blocks of diagrams.

Feature List

Right now, we list some points above that are mandatory for this project, like Build a system that visualizes information on a general level, Visualize first to identify interesting features/observations, Featuring the engineering, Using fillering to combine graphs into one, and other numbers and statistics by storytelling. We hope we could raise more detail into our problem statements in the future.

<u>The interaction element:</u>

- Tooltips on each chart to show details
- Selection brush to zoom in the heatmap by time
- A filter for all pattern chart to filter out desired car type
- Scale-changeable timeline chart for individual car-id

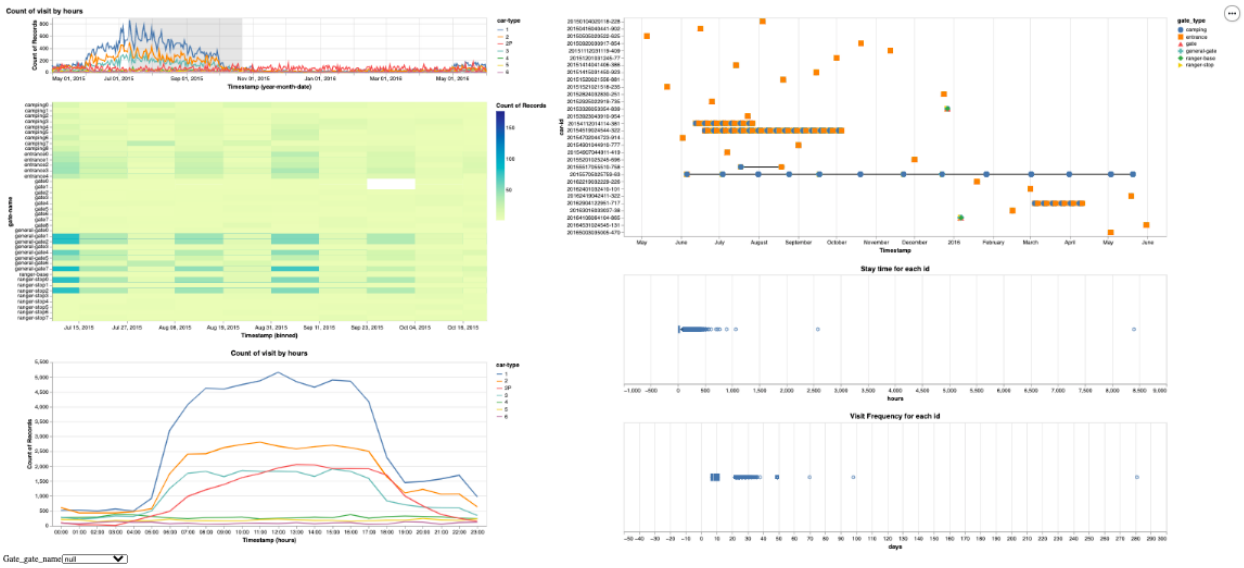
Description of team role:

In phase1, and phase2, each of us raised at least two tasks in this project. Based on the questions, we design the data visualization together. Zeyu wrote the code; based on the data visualizations, Kan and Irene specified the problem statements and changed some variables.

In the Final phase, we are trying to split the coding work evenly and then collaborate them to finalize the system.

FINAL PHASE: Findings and conclusions:

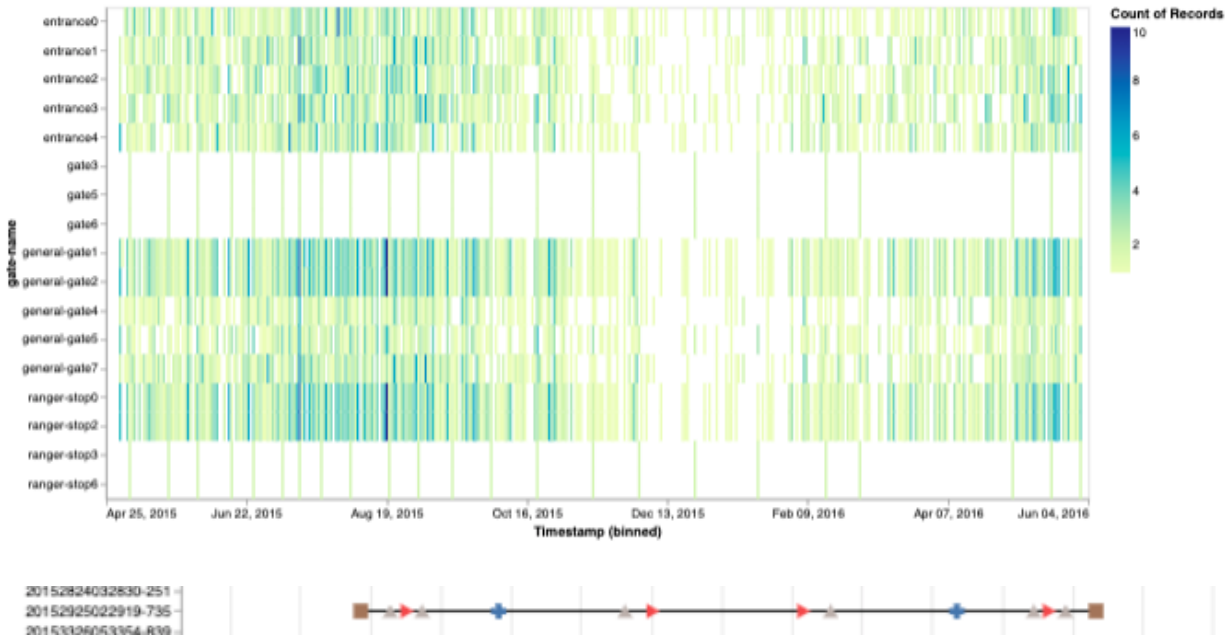
1. **<u>The final system:</u>**



2. **<u>Approach and Data analysis process:</u>** The data cleaning and exploratory visualization is mostly done with python. We have separate sets of the clean data, each of which is used for a separate point of analysis. We carried out the process on our own and shared the results together. We selected the datasets that contain most interesting features out of all the features we could create out of the existing features. Then, the in-depth exploration is done with Tableau. We created different dashboards to analyze the data. The major patterns are noted down for the design of the system. We selectively picked the features that not only manifest the key findings but also are connected to each other. The goal is to keep the design simple but informative. We then split up the work and each person worked on a few graphs before we put them all together. The interaction and tooltip is added at the end and small adjustments are made to better fit the system. For example, initially, we thought we should display the travel timeline of all car-ids, but

in that way, the system will be tremendously overwhelming so we decided to only include the itinerary of those cars of interest.

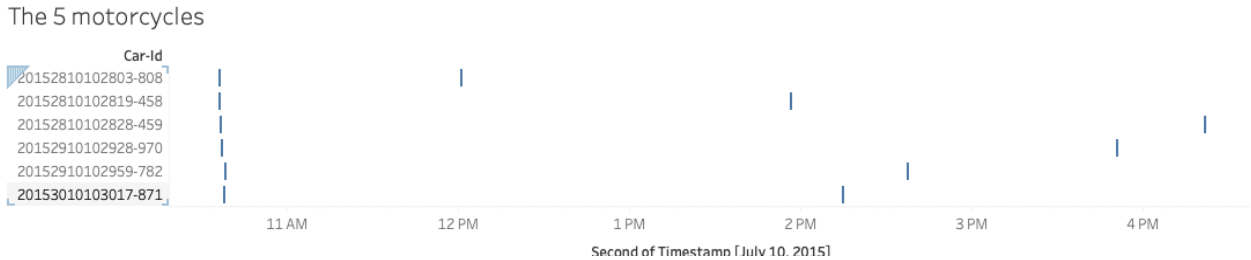
3. **Findings:** In the end, we find that two out of our 3 initial hypotheses match up to our analysis, the second and the third. For the general claim in the second hypothesis, we have been able to further detail and locate specifically the patterns that might be responsible for the decline in bird population in the preserve. The details of all the odd behaviors observed throughout the whole analysis process is documented in the answer sheet. Some key findings that might be the cause of our problem are:
    - There are some vehicles type 4 that have been going past the gates and ranger-stop areas that are not commonly seen among such vehicles in the preserve during early mornings. Notably, the route that these vehicles take is almost the same every time. This occurrence goes on for a few months. It's reasonable to believe there is a motivation behind this pattern, which we believe might be responsible for the decline in bird population.



- The vehicle number 63 has been in the preserve for over a few months already. They have changed location multiple times. Each time they relocate, they go past a few different gates, and stay at the new location for a period (approximately a month, sometimes a quarter of a month), before moving to the next location. We don't know if

they have been granted the access by the preserve to stay for such a long time for some reason (filming a documentary, etc) or they have been able to stay unrecognized to us.

- Vehicles 381 and 322 have been visiting the preserve for a lot of times over a few months already. Each time they stay over a few days and go on a very similar itinerary. Even though this can be seen as frequent visitors, it's still advisable to take a closer look into such behaviors.
- There are 5 vehicles type 1 (motorcycle) going into the gate area at the same time. Even though we think that they might not be the culprit since this occurs only once, we still think it's reasonable to tighten up the control against such trespassing behaviors because they can inadvertently impose harm on wildlife if they enter such a specialized area.



- There are a few vehicles that visit the preserve very shortly, without stopping at any sites. We conjecture such vehicles might be finding a shortcut by going through the preserve. These behaviors may harm the cultivated habitat, the bird population included, in that they will bring a huge amount of emission into the preserve, which should be brought under control as well.

4. Conclusion:

The pattern of life helps us identify what are the “normal behaviors", and distinguish between them and the “odd" ones. Overall, we find that the cause of the problem may not be because of the large number of visitors in a season, but more because of the eccentric individual patterns that exist over a longer period of time. Those are the patterns that are hidden if we just look at the daily levels, but viewing it over a period of time, we can see suspicious cycles of things. The next step will be focusing on these behaviors to address if they have any motivation behind such activities and find a solution to it. For now, we think that tightening up the security as well as limiting the number of cars (non-visitors) into the preserve might have a positive impact on

the bird population and the wildlife. Further statistical testing is required to confirm the effect as well as the usefulness of the system.