Report MC2

Autoren: Jan Klaassen, Johannes Paul

Einleitung

In dem Dokument wird die VAST Mini-Challenge 2 bearbeitet. Bei der Aufgabe handelt es sich um die Verarbeitung von Kommunikationsdaten innerhalb eines Freizeitparks.
Der Zeitraum der erfassten Daten erstreckt sich über ein Wochenende (Fr.-So.) und soll helfen das Verhalten der Besucher zu analysieren. Außerdem sollen die Daten genutzt werden, um Informationen über ein am Wochenende aufgetretenes Verbrechen zu gewinnen.
Die Analyse der Daten erfolgt mit Hilfe der Programmiersprache Python. Dabei wurden die Daten eingelesen,aufbereitet und visualisiert.

Exercise 1

Identify those IDs that stand out for their large volumes of communication. For each of these IDs
A) Characterize the communication data you see
B) Based on these patterns, what do you hypothesize about these ID's?

Lösungsansatz:

In dieser Aufgabe wird auf die hervorstechenden IDs bezüglich der eingehenden- und getätigten Anrufe eingegangen und es werden Thesen aufgestellt, wieso diese Daten hervorstechen.
A) In Tabelle 1(Tabelle getätigte Anrufe?!?) der getätigten Anrufe lassen sich in den Daten zwei extreme Ausreißer ausmachen:

- ID 1278894 mit 190360 Anrufen
- ID 839736 mit 60812 Anrufen

from	calls	unique callers	friday calls	saturday calls	sunday calls
1278894	190360	2521	38658	70143	81559
839736	60812	8720	5914	10224	44674
1045021	3807	376	1429	1570	808
1116329	3746	454	1477	1289	980
1749109	3708	416	1204	1520	984

from	coaster	entry	kiddie	tundra	wet
1278894	0.0	190360.0	0.0	0.0	0.0
839736	0.0	60812.0	0.0	0.0	0.0
1045021	312.0	349.0	183.0	710.0	2253.0
1116329	711.0	259.0	61.0	955.0	1760.0
1749109	634.0	278.0	254.0	1178.0	1364.0

Tabelle 1: Aggregierung nach den fünf häufigsten Anrufer('from')

In der Tabelle (Tabelle getätigte Anrufe) ist zu sehen, dass die beiden Ausreißer ihre Anrufe jeweils nur aus dem Bereich ‘Entry Corridor’ tätigen.

Bei der Anzahl der eingehenden Anrufe pro ID (Tabelle eingehende Anrufe?!?) lassen sich drei verschiedene Ausreißer beobachten:

- ID 1278894 mit 189894 eingehenden Anrufen
- ID external mit 62077 eingehenden Anrufen
- ID 839736 mit 60818 eingehenden Anrufen

to	calls	unique callers	friday calls	saturday calls	sunday calls	\
1278894	189894	2521	38540	70001	81353	
external	62077	9265	11302	21081	29694	
839736	60818	8720	5914	10223	44681	
171002	3270	502	1241	1252	777	
1116329	3153	492	1279	1020	854	

to	coaster	entry	kiddie	tundra	wet
1278894	25303.0	28775.0	19483.0	41060.0	75273.0
external	6968.0	8624.0	5303.0	12795.0	28387.0
839736	7741.0	4161.0	2621.0	6463.0	39832.0
171002	324.0	506.0	299.0	662.0	1479.0
1116329	320.0	461.0	358.0	729.0	1285.0

Tabelle 2: Aggregierung nach den fünf häufigsten angerufenen IDs(‘to’)

B) Bei den beiden Ausreißern (ID ‘1278894’ und ID ‘839736’) lässt sich vermuten, dass es sich um Servicenummern des Funparks handelt. Diese Vermutung begründet sich durch die folgenden zwei Grundlagen:

- Beide Nummern weisen eine extrem hohe Anzahl an Kommunikation, sowohl durchgeführte, als auch eingehende Anrufe, auf.
- Die Anrufe Beider IDs kommen ausschließlich aus dem Eingangsbereich (‘Entry Corridor’)

Zur Veranschaulichung und Analyse des Kommunikationsverhalten von ID ‘1278894’ dient Abbildung 1 (‘Calls in each hour for ID=1278894’). Hier ist das Anrufverhalten der ID über das gesamte Wochenende zu sehen. Dabei wird jeweils nacheinander die stündliche Kommunikation der verschiedenen Tage aufgezeigt.
Die Grafik verdeutlicht, dass die ID in einem 2 Stunden Rhythmus Nachrichten an alle IDs verschickt. Der gleichmäßige Rhythmus lässt eine Service-Benachrichtigung des Parks vermuten, die an alle Besucher des Parks verschickt wird.
Die hohe Anzahl der Anrufe der ID ‘external’ ist dadurch zu begründen, dass es sich bei dieser ID um die Kommunikation von IDs innerhalb des Parks mit der Außenwelt handelt.

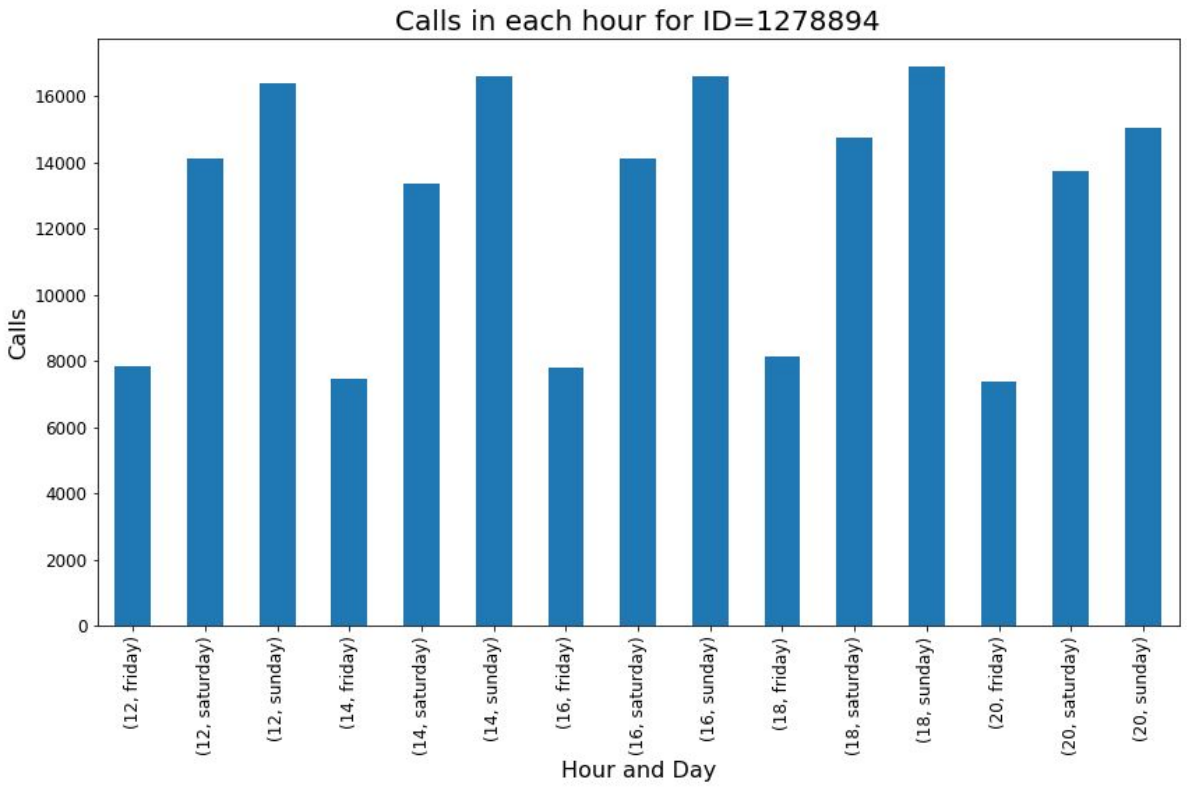

Abbildung 1 'Calls in each hour for ID= 1278894'

Exercise 2

Describe up to 10 communication patterns in the data. Characterize who is communicating with whom, when and where.

Lösungsansatz

Um einen Überblick über die Häufigkeit der durchgeführten Anrufe pro ID zu geben wurde dient Abbildung 2. Dabei wird gezeigt, dass sich der Großteil der getätigten Anrufe pro ID auf den Bereich von 0-3000 beschränkt. Dabei befinden sich der Hauptanteil in dem Bereich von 0-200 (der erste Balken in Abbildung 2).

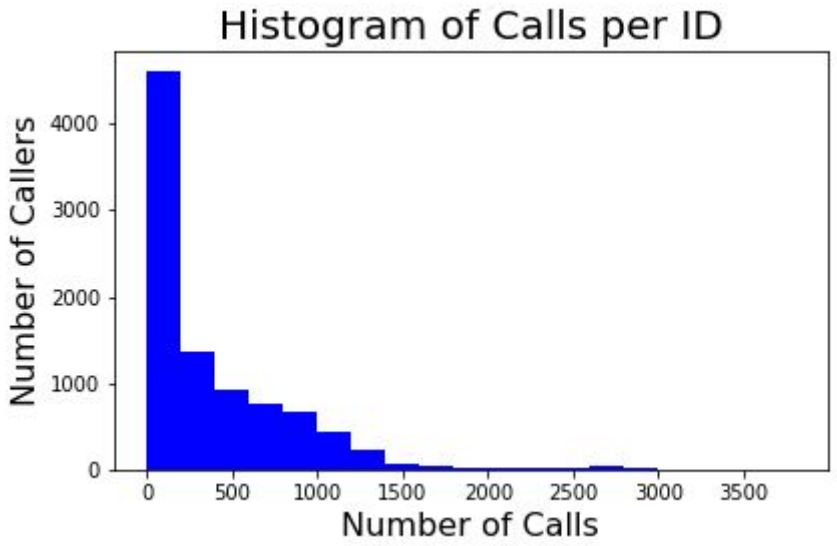

Abbildung 2 'Histogram of Calls per ID'

Um einen generellen Überblick über die getätigten Anrufe im Verlaufe des Wochenendes zu bekommen dienen die folgenden Grafiken:

- Abbildung 3 'Barplot of calls per hour on friday'
- Abbildung 4 'Barplot of calls per hour on saturday'
- Abbildung 5 'Barplot of calls per hour on sunday'

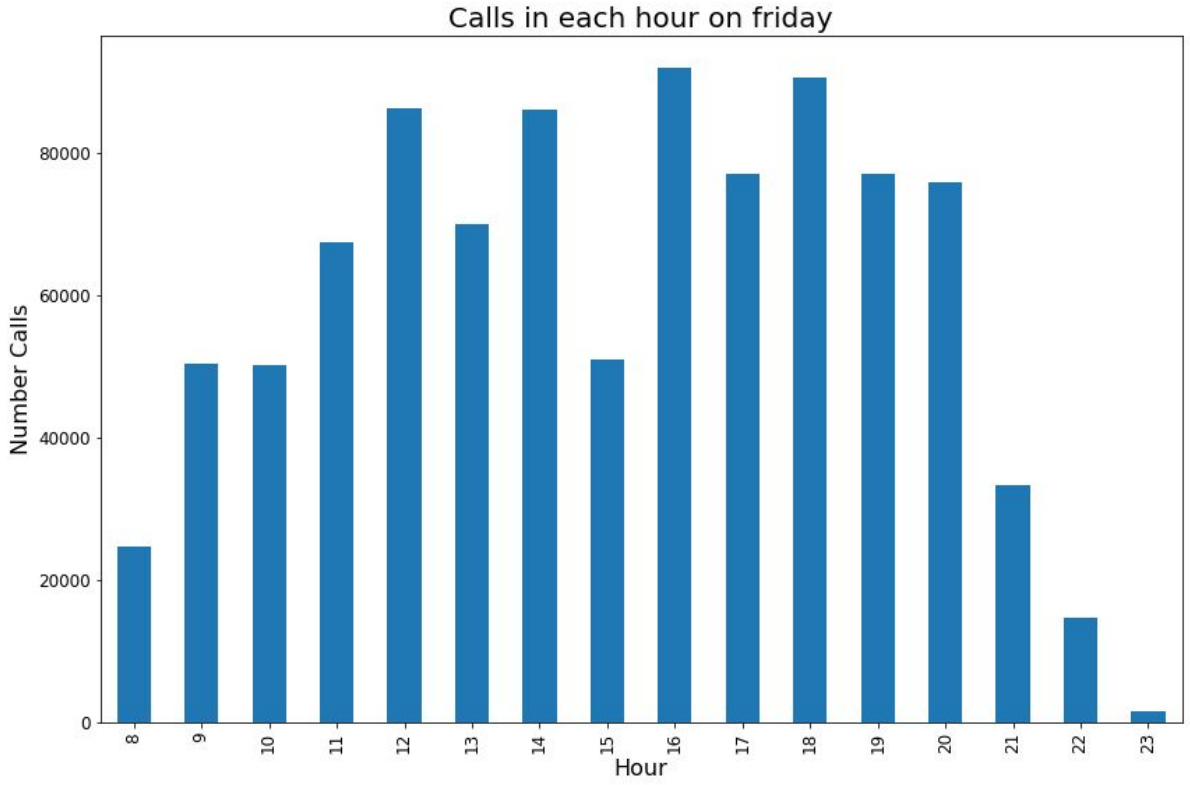

Abbildung 3 'Barplot of calls per hour on friday'

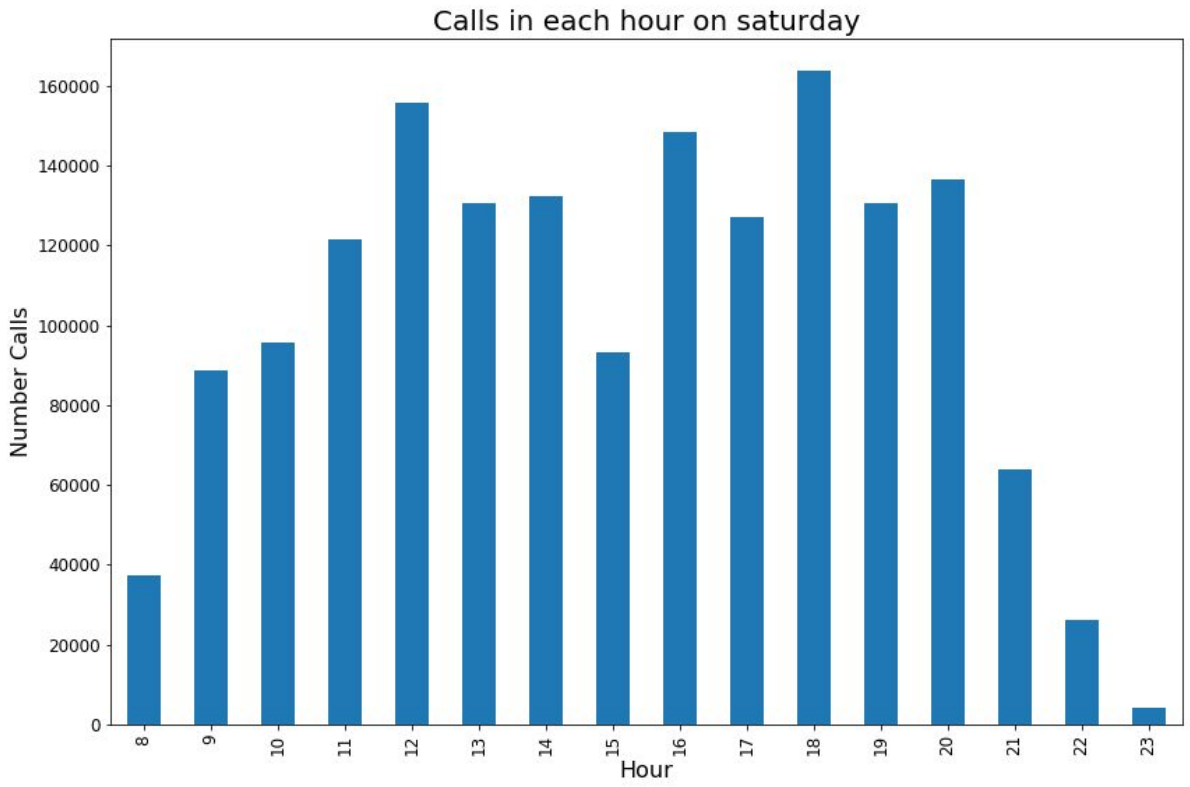

Abbildung 4 'Calls in each hour on saturday'

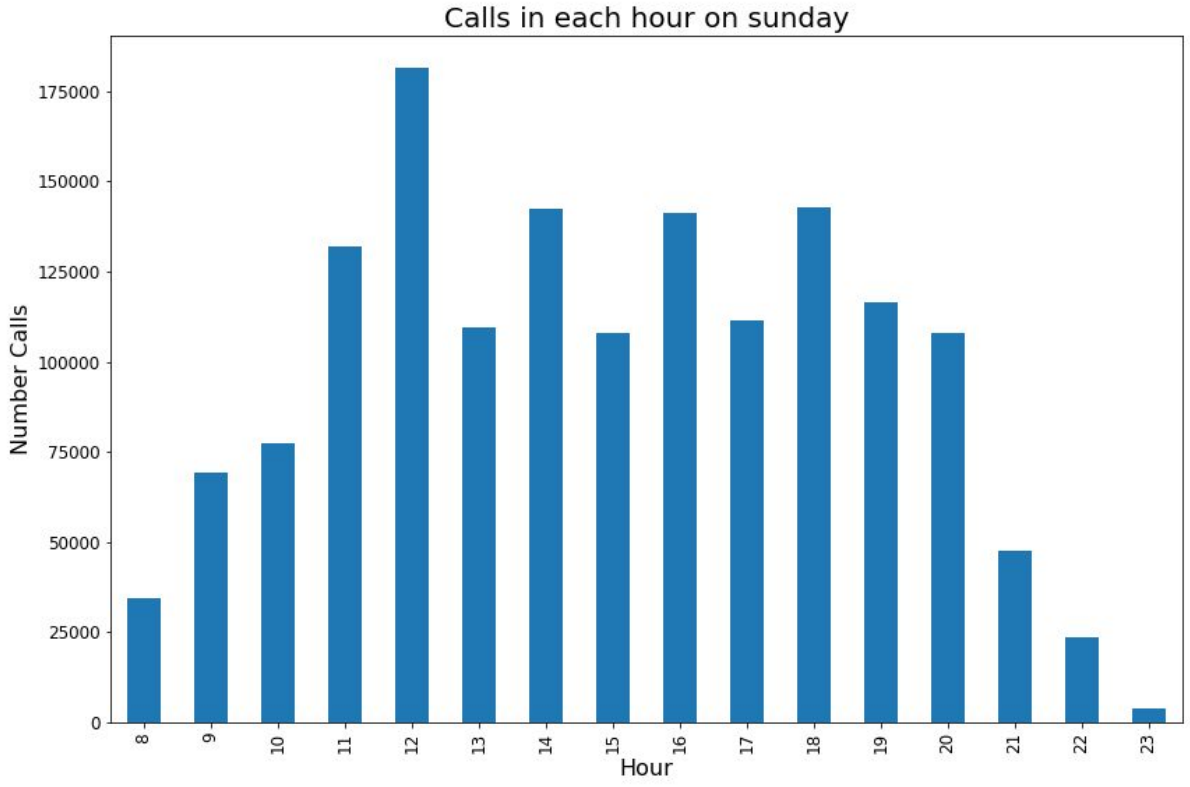

Abbildung 5 'Calls in each hour on sunday'

In den Grafiken (Abbildung 3-5) wird jeweils der Überblick über die durchgeführten Anrufe im ein Stunden Takt aufgezeigt.
Dabei ist zu erkennen, dass die Kommunikation an allen Tagen am Morgen und Abend erst gering ist, da die Besucherzahlen zu diesen Zeiten geringer ist als am Rest des Tages.
Bei dem Vergleich der Anrufe der verschiedenen Tage wird deutlich, dass am Mittag / Nachmittag meist eine regelmäßige Schwankung in einem 2h Takt auftritt.
Die Schwankung lässt sich auf die Benachrichtigungen der ID ‘1278894’ zurückführen (siehe Abbildung 1), welche in einem Rhythmus von 2h Nachrichten versendet.
Die meiste Kommunikation des Wochenendes findet am Sonntag zwischen 12 und 13 Uhr mit über 175000 Anrufen statt. Im Gegensatz zu Samstag und Freitag ist an diesem Tag ein hervorstechendes Maximum zu erkennen.

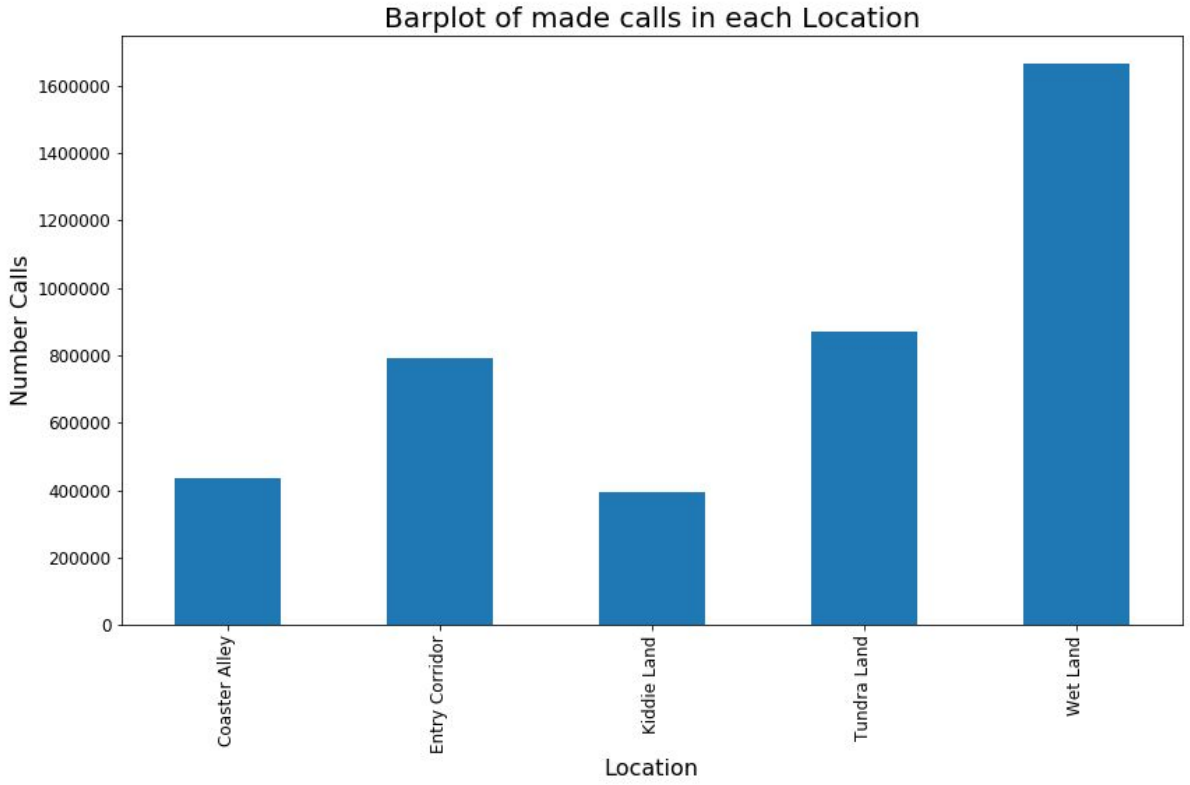

Abbildung 6 ‘Barplot of made calls in each Location’

Ein Vergleich der getätigten Anrufe der Unterschiedlichen Bereiche des Parks ist durch Abbildung 6 (‘Barplot of made calls in each Location’) dargestellt. In der Abbildung ist zu sehen, dass in den Bereichen ‘Coaster Alley’ und ‘Kiddie Land’ am wenigsten Anrufe durchgeführt werden. Dies ist möglicherweise darauf zurückzuführen, dass es sich hierbei um aktive Bereiche handelt. In diesen Bereichen ist rückt die Kommunikation durch Aktivitäten eher in den Hintergrund.
In dem Eingangsbereich ‘Entry Corridor’ und dem Bereich ‘Tundra Land’ die Anzahl der getätigten Anrufe doppelt so hoch wie in den zuvor genannten Bereichen.
Der Bereich ‘Wet Land’ weist die höchste Anzahl an getätigten Anrufen auf. In diesem Bereich wurden doppelt so viele Anrufe getätigt wie in den Bereichen ‘Entry Corridor’ und ‘Tundra Land’. Dies lässt sich darauf zurückführen, dass an dem Wochenende ein Verbrechen in dem Pavilion stattgefunden hat. Dieser besitzt nur einen Zugang, welcher von

dem ‘Wet Land’ ausgeht. Vermutlich ist die Anzahl der Anrufe hier so hoch, da die Besucher das Verbrechen an dem Pavilion gemerkt haben und dies melden wollten. Bei dem Melden des Verbrechens haben die Personen sich auf dem ‘Wet Land’ befunden.

Exercise 3

From this data, can you hypothesize when the crime was discovered? Describe your rationale.

Lösungsansatz

Wenn ein Verbrechen von den Besuchern des Parks festgestellt wird, wird dies bei der Servicenummer des Parks gemeldet. Um den Zeitpunkt festzustellen, wann das Verbrechen bemerkt wurde, muss ein Ausreißer bei den eingehenden Anrufen bei einer der in Aufgabe 1 festgestellten Servicenummern des Parks gesucht werden.

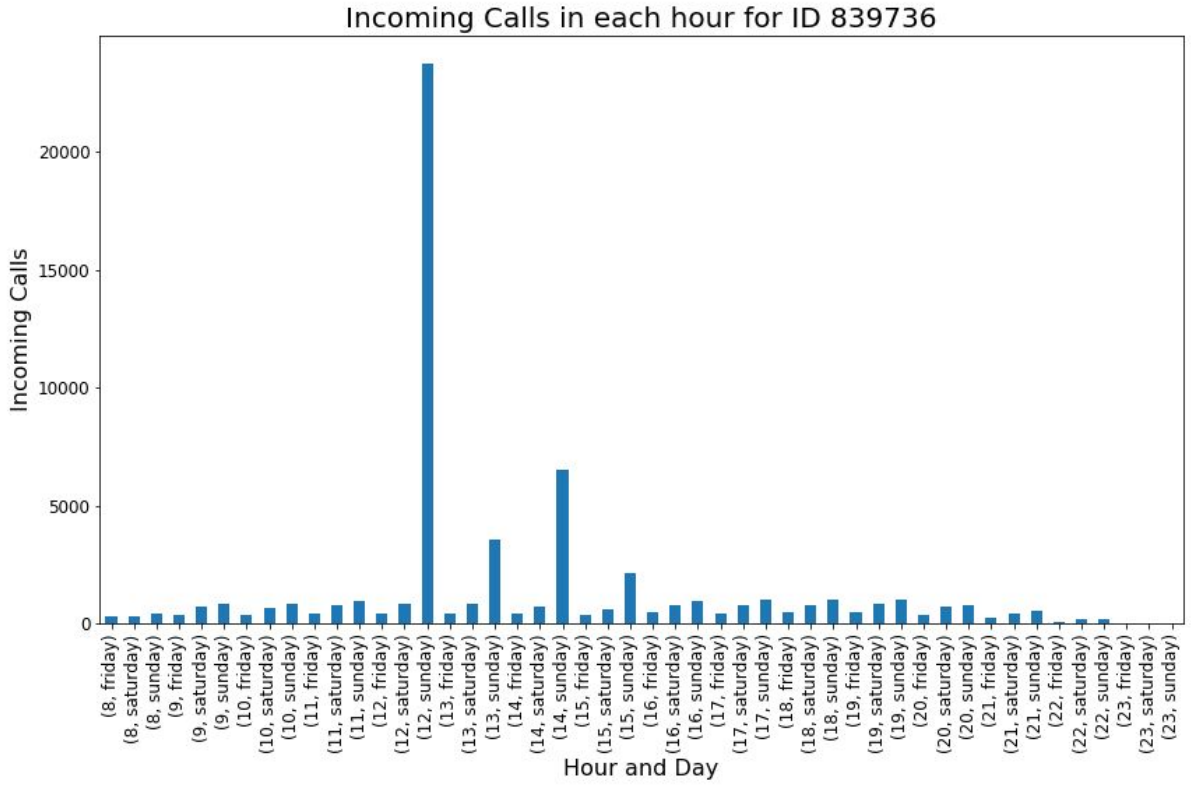

Abbildung 7 ‘Incoming Calls in each hour for ID 839736’

Die Abbildung 7 (‘Incoming Calls in each hour for ID 839736’) zeigt die Anrufe, die jeweils zu den verschiedenen Stunden der jeweiligen Tage bei der ID ‘839736’ eingegangen sind. Dabei ist zu erkennen, dass es zu dem Zeitpunkt zwischen 12 und 13 Uhr am Sonntag einen signifikanten Ausreißer gibt. Dies lässt somit darauf schließen, dass zu diesem Zeitpunkt das Verbrechen bemerkt wurde. Die Anschließenden 2 Stunden (13-15 Uhr) weisen auch eine erhöhte Anzahl eingehender Anrufe aus. Dies kommt vermutlich dadurch zustande, dass der Ort des Verbrechens vielleicht abgesperrt wurde und Leute sich erkundigen wollten, weshalb dies so ist.