Análisis del Sistema de **Transporte Urbano de Pasajeros** de la Ciudad de Córdoba

Presentación
Final

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones
FAMAF-2022

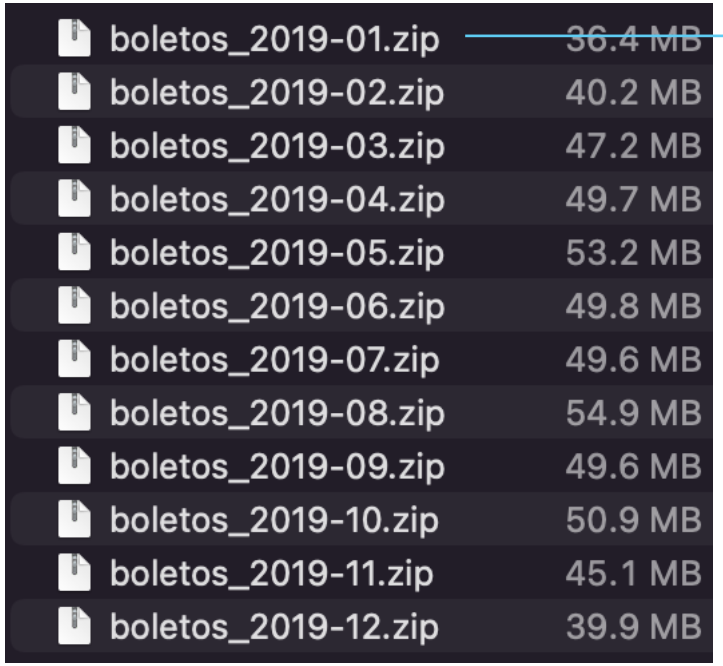
Mentor
Néstor Grión

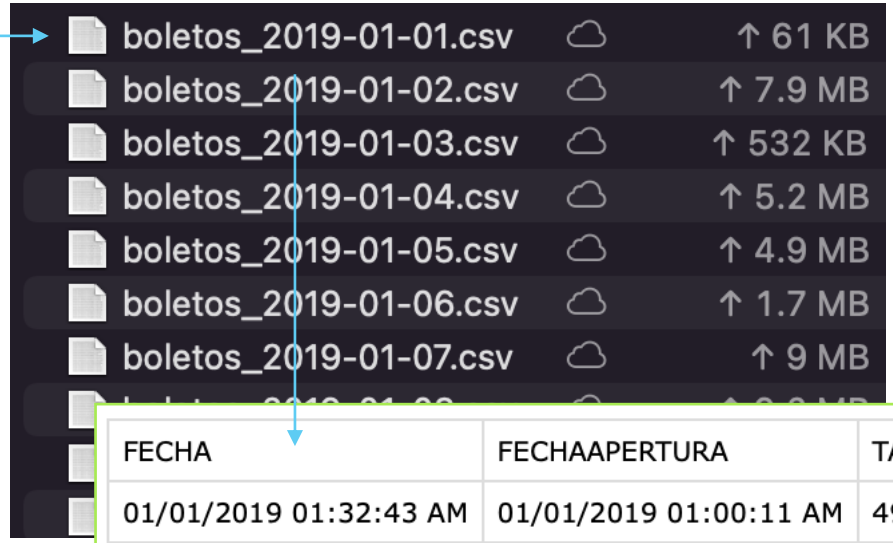
Grupo 1
Juan A. Fraire
Ignacio Villagra
Torcomian
~~Esteban Kocian~~

Agenda

- Datos
- Metodología
- Análisis
  - **P1** Analisis de estacionalidad
  - **P2** Analisis de uso de tarjetas
  - **P3** Aprendizaje supervisado
  - **P4** Aprendizaje no supervisado
- Cierre

Datos





Descomprimido en su totalidad, el dataset ocupa **múltiples GB en disco/RAM** lo que hace muy difícil de trabajar con todos los datos del 2019

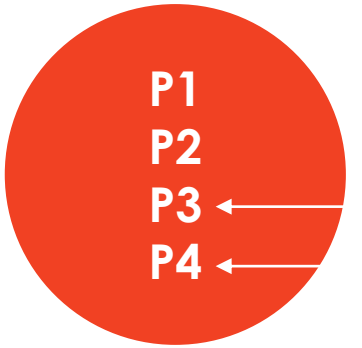
FECHA	FECHAAPERTURA	TARJETA	CORREDOR	LINEA	SENTIDO
01/01/2019 01:32:43 AM	01/01/2019 01:00:11 AM	4992639	Cor 3 Rojo	L35	Ida
01/01/2019 02:26:11 AM	01/01/2019 01:00:11 AM	5489022	Cor 3 Rojo	L35	Vuelta
01/01/2019 02:26:14 AM	01/01/2019 01:00:11 AM	5489022	Cor 3 Rojo	L35	Vuelta
01/01/2019 02:32:27 AM	01/01/2019 01:59:31 AM	5317539	Cor 1 Naranja	L12	Vuelta
01/01/2019 02:57:48 AM	01/01/2019 01:59:31 AM	4451141	Cor 1 Naranja	L12	Vuelta
01/01/2019 03:03:55 AM	01/01/2019 02:41:18 AM	6002199	Cor 1 Naranja	L14	Ida
01/01/2019 03:23:17 AM	01/01/2019 02:41:18 AM	5668125	Cor 1 Naranja	L14	Ida
01/01/2019 03:50:55 AM	01/01/2019 02:41:18 AM	3841688	Cor 1 Naranja	L14	Vuelta

Un archivo **.zip** comprimido por cada mes → contiene un archivo **.csv** por cada día → contiene una **línea por cada boleto** usado

Metodología

- **Tractabilidad**
  - Priorizamos el análisis integral a lo largo del año 2019
- **Estrategia**
  - 1) **Pre-Procesador** (Python)
    - Genera datasets más pequeños en basa al dataset maestro
    - Agrega y agrupa datos en función del análisis
    - Toma hasta **5 horas** de cómputo, pero se ejecuta una sola vez
  - 2) **Post-Procesador** (Python)
    - Carga los datos de los datasets pequeños
    - Realiza cómputos estadístico y genera gráficas en matplotlib
    - Ejecución ágil

Dejamos el análisis por recorrido para enfocarnos en el uso de tarjetas



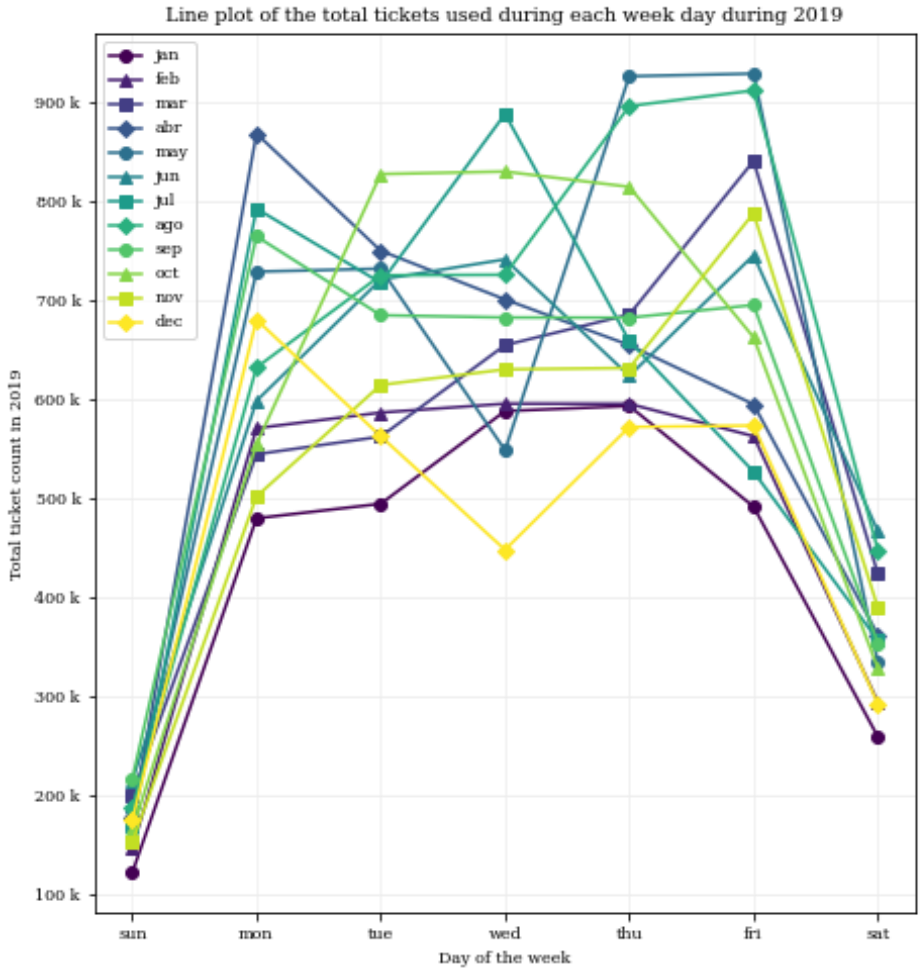
P1 Análisis de Estacionalidades

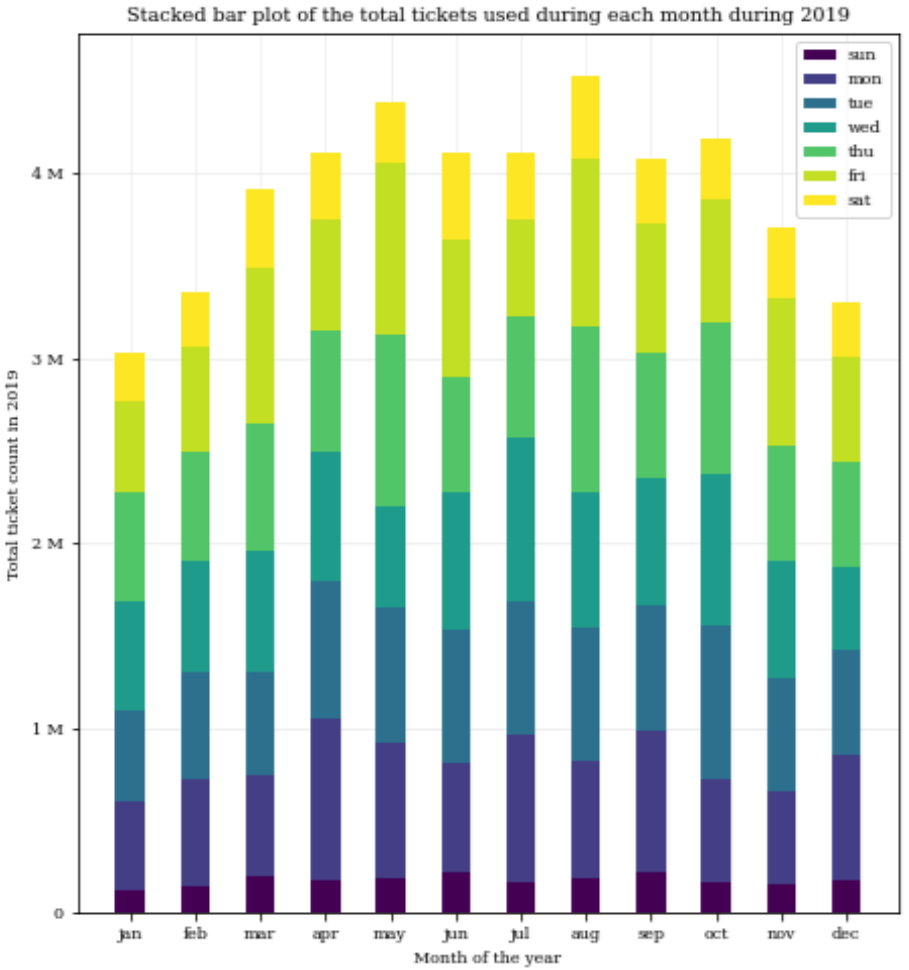
month	acc-month	acc-sun	acc-mon	acc-tue	acc-wed	acc-thu	acc-fri	acc-sat
1	3027469.0	121409.0	479719.0	494417.0	588208.0	593321.0	491533.0	258862.0
2	3350995.0	145449.0	570845.0	586678.0	595639.0	595502.0	562766.0	294116.0
3	3911586.0	199273.0	544477.0	562360.0	654852.0	685807.0	841094.0	423723.0
4	4109047.0	176966.0	868809.0	750281.0	701089.0	654868.0	595041.0	361993.0
5	4386188.0	185486.0	729178.0	732199.0	548657.0	926816.0	929475.0	334377.0
6	4111966.0	214256.0	598524.0	721195.0	741616.0	623840.0	745737.0	466798.0
7	4110114.0	168547.0	792930.0	717822.0	888671.0	658984.0	526801.0	356359.0
8	4527237.0	186457.0	632277.0	725505.0	726027.0	896683.0	912465.0	447823.0
9	4080865.0	216302.0	764880.0	685130.0	682847.0	682629.0	695455.0	353622.0
10	4183047.0	164173.0	554739.0	827908.0	830442.0	814950.0	662842.0	327993.0
11	3708146.0	151868.0	501840.0	614586.0	630348.0	631735.0	788253.0	389516.0
12	3303449.0	174522.0	680657.0	562609.0	447890.0	572112.0	573546.0	292113.0
Total	46810109.0	2104708.0	7718875.0	7980690.0	8036286.0	8337247.0	8325008.0	4307295.0
%	100.0	4.5	16.5	17.0	17.2	17.8	17.8	9.2
min	3027469.0	121409.0	479719.0	494417.0	447890.0	572112.0	491533.0	258862.0
max	4527237.0	216302.0	868809.0	827908.0	888671.0	926816.0	929475.0	466798.0

	month	acc-month	%
0	1	3027469	6.5
1	2	3350995	7.2
2	3	3911586	8.4
3	4	4109047	8.8
4	5	4386188	9.4
5	6	4111966	8.8
6	7	4110114	8.8
7	8	4527237	9.7
8	9	4080865	8.7
9	10	4183047	8.9
10	11	3708146	7.9
11	12	3303449	7.1
12	Total	46810109	100.0

Datos para la **totalidad** del dataset (**201**

P1 Análisis de Estacionalidades

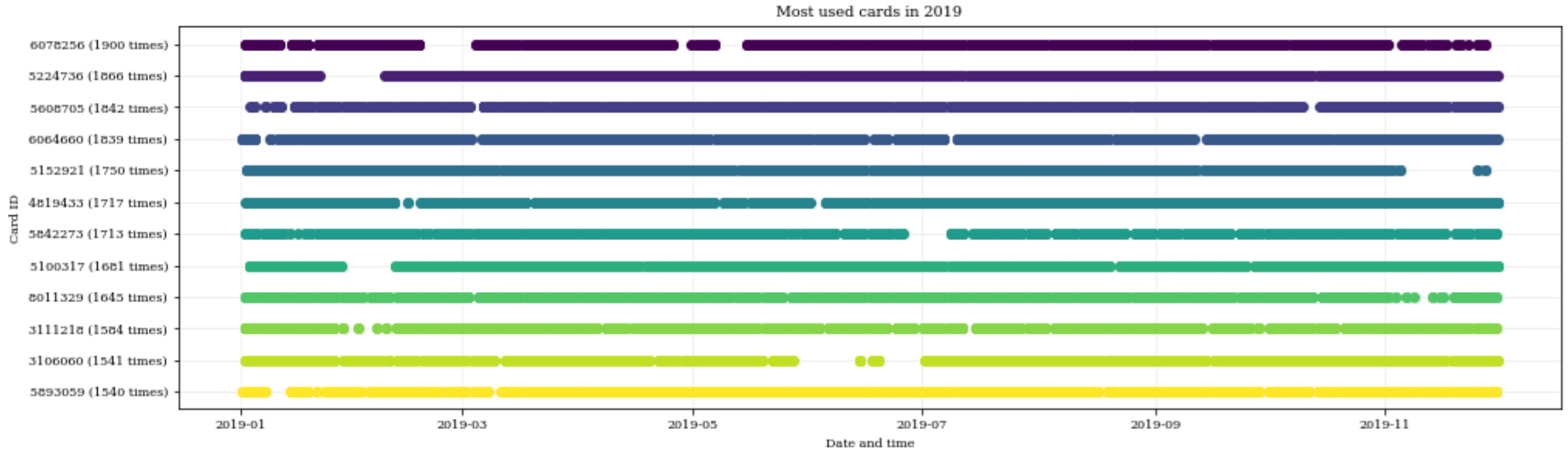


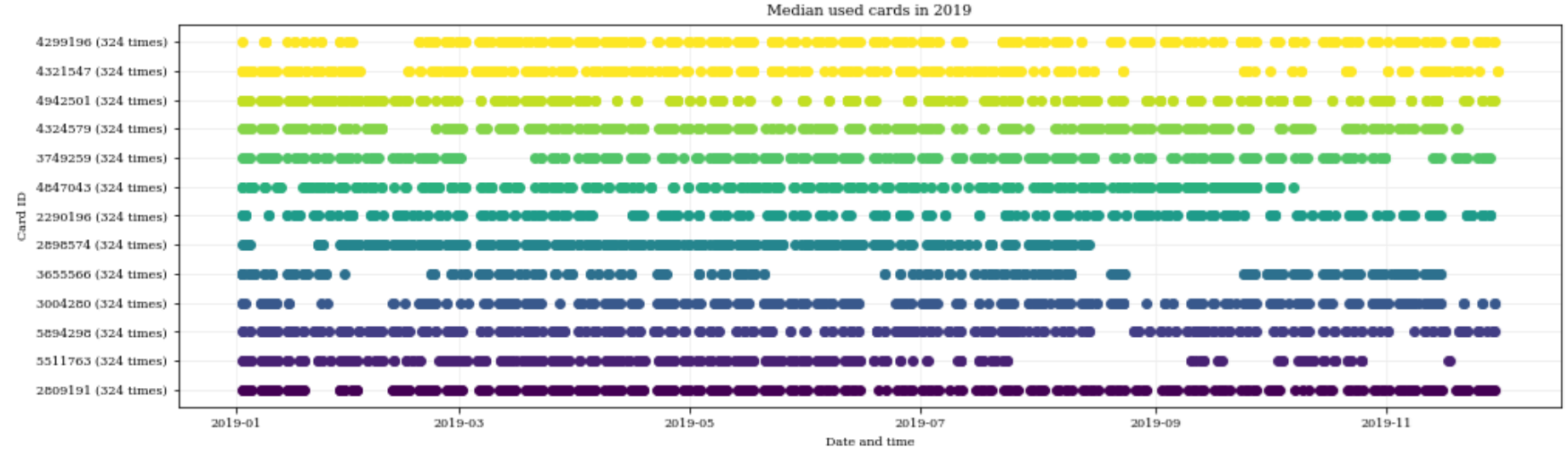


P1 Análisis de Estacionalidades

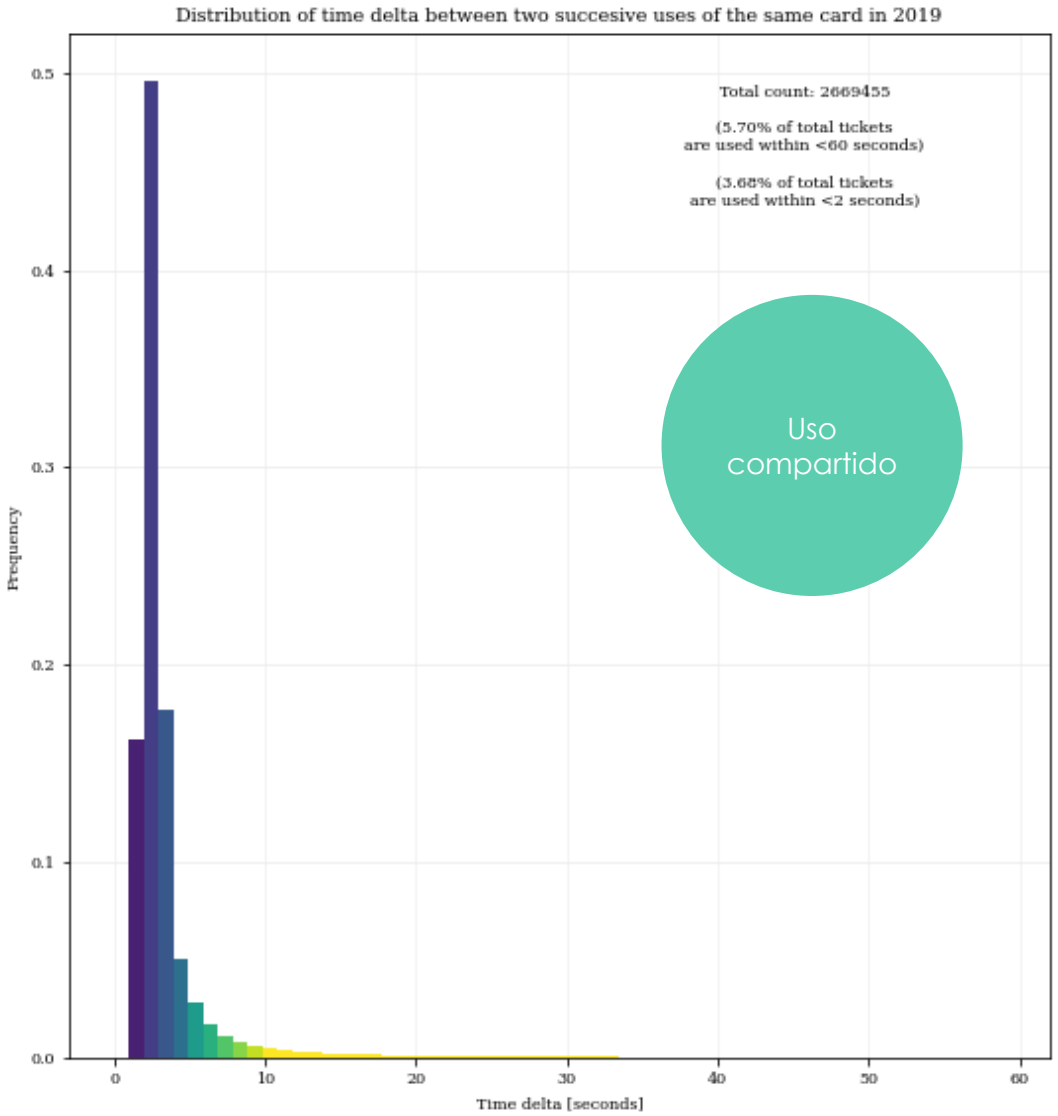
- **Generalidades**
  - Total: el total de boletos en el 2019 es de 46810109 (46.8 millones de boletos)
- **Estacionalidad Diaria**
  - Días laborales: 16.5% (**Lun**), 17.1% (**Mar**), 17.0% (**Mie**), 17.7% (**Jue**) y 17.9% (**Vie**)
  - Días no laborales: 9.2% (**Sab**) y 4.5% (**Dom**)
  - Mínimos: **Dom**, **Lun**, **Mar**, **Mie**, **Vie** y **Sab** en Enero, **Jue** en Diciembre
  - Máximos: **Dom** Sept, **Lun** Abril, **Mar** Oct, **Mie** Jul, **Jue** May, **Vie** May, y **Sab** Jun
- **Estacionalidad Mensual**
  - Menor participación en el proporcional de corte de boletos: **Enero** con 6.5%
  - Mayor participación en el proporcional de corte de boletos: **Agosto** con 9.7% (seguido por **Mayo** con un 9.4% del total anual)

P2 Análisis del Uso de Tarjetas

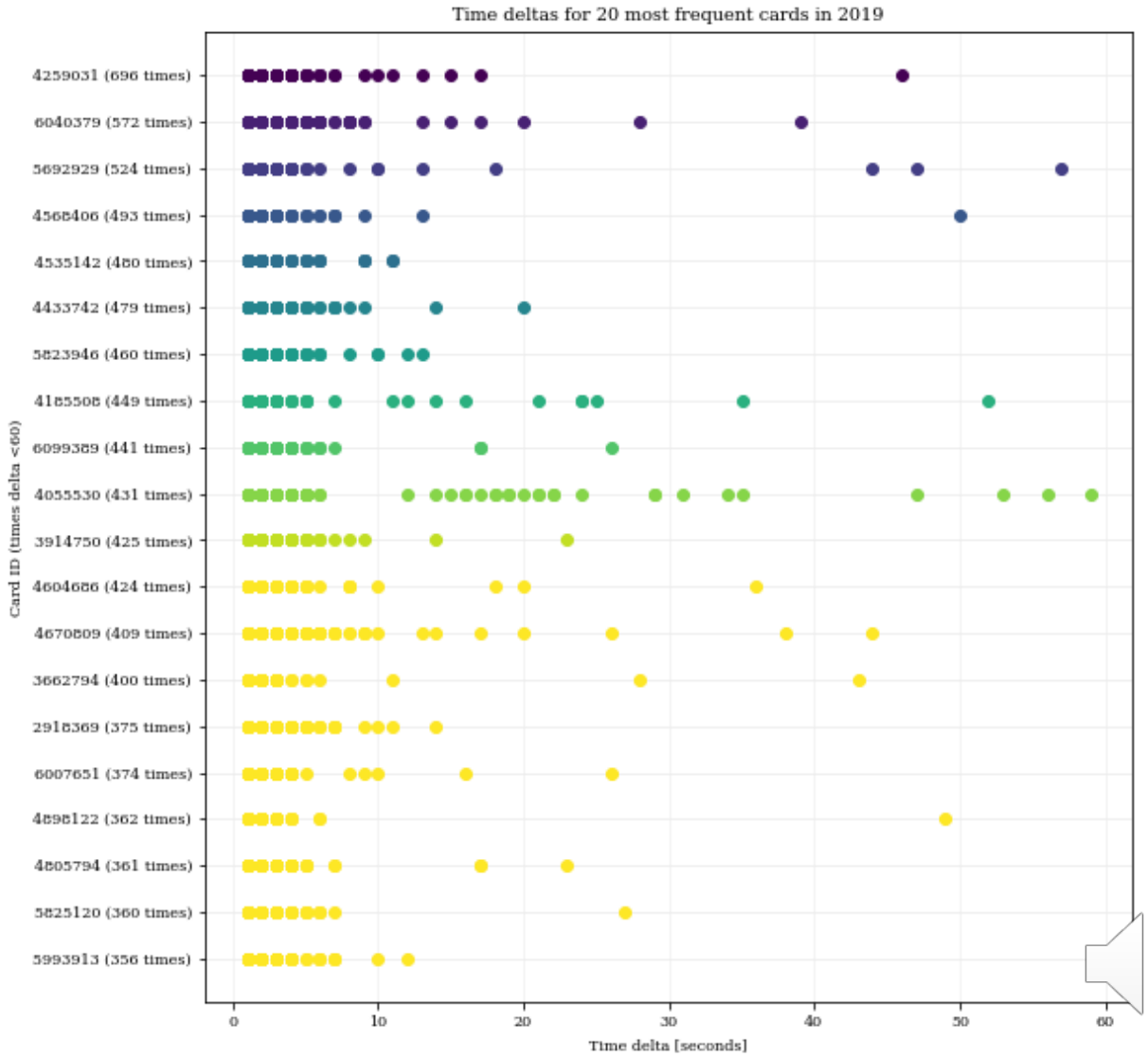




P2 Análisis del Uso de Tarjetas



Uso compartido



P2 Análisis del Uso de Tarjetas

- **Generalidades**
  - Máximo: la tarjeta más usada registra **1900 boletos** en el año
  - Media: la media de la utilización de tarjetas es de **324 boletos**
  - Total: el total de tarjetas usadas asciende a **118691 tarjetas**
- **Uso compartido** (dos usos dentro de 60s)
  - Total boletos: el total de boletos compartidos es de 2.6 mil. (**5.7%** del total)
  - Menor a 2 segundos: **3.68%** son boletos compartidos dentro los 2 segundos
  - Mayor uso: la tarjeta más compartida se compartió **696 veces**
  - Total tarjetas: el total de tarjetas que al menos han compartido un boleto asciende a **117793 tarjetas** (**99%** del total)

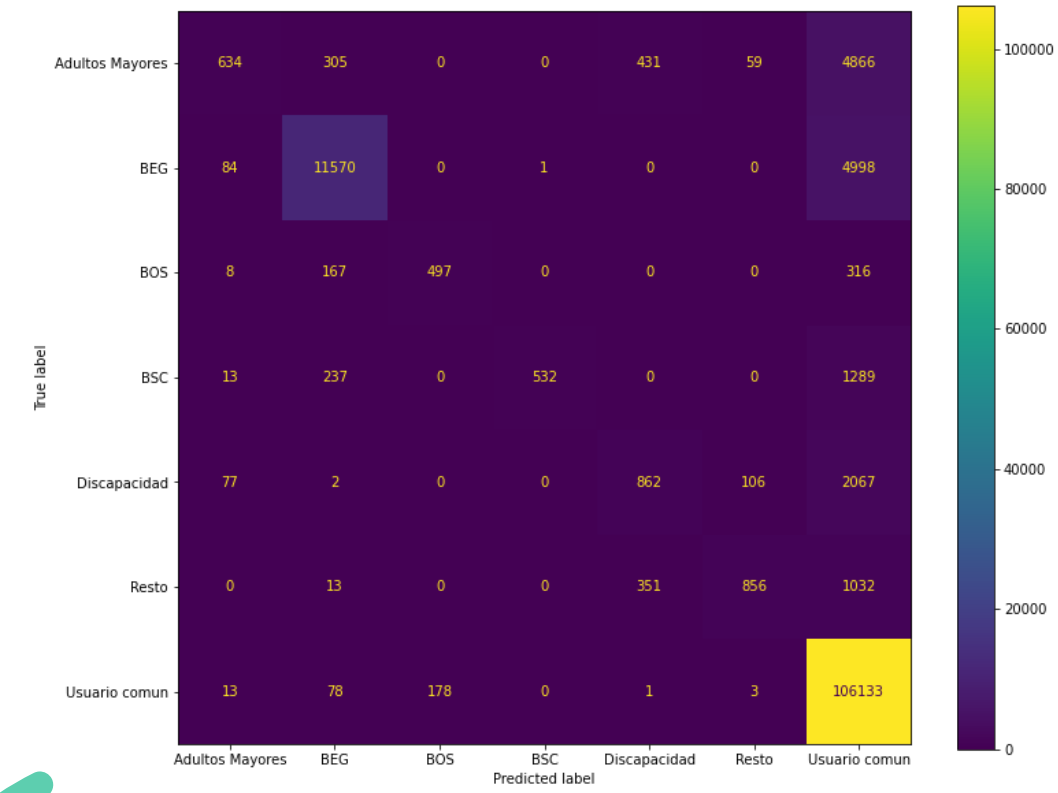
P3 Aprendizaje Supervisado

- **Generalidades**
  - Ingeniería de características:
    - **Dataset**: Una fila por tarjeta, contratos y variables relevantes
    - **Preparación**: Tarjetas con múltiples contactos removidas del dataset
    - **Variables**: cantidad de viajes, cantidad de líneas distintas de colectivo que utiliza el pasajero, línea de mayor frecuencia, cantidad de días distintos de viaje, primer y último día de viaje, primera hora, última hora y hora promedio de viaje.
  - Algoritmos de aprendizaje supervisado:
    - **Decision Tree Classifier:** Accuracy test: 87.88%
    - **Random Forest Classifier:** Accuracy test: 84.12%
  - Evaluación sin usuarios 'comunes': e
    - **Decision Tree Classifier:** Accuracy test: 75.26%
    - **Random Forest Classifier:** Accuracy test: 70.37%

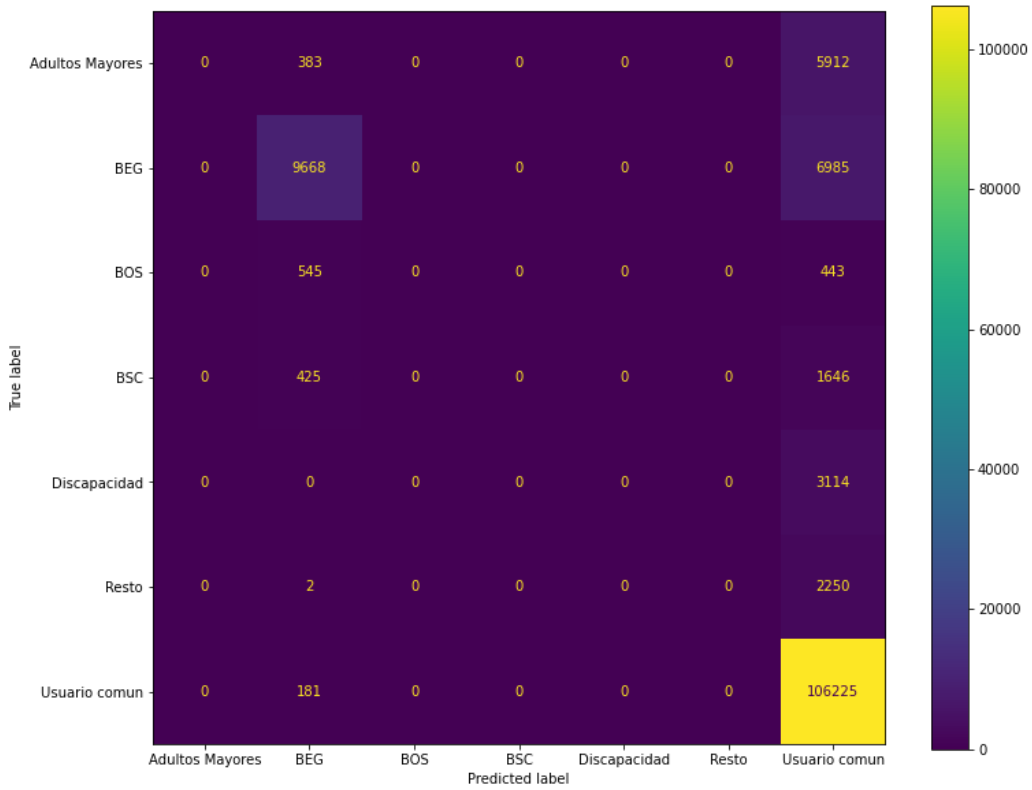
Random Forest: solo se obtuvieron métricas elevadas de *precision*, *recall* y *f1* de los contratos *Común* y *BEG* (los dos de mayor frecuencia)

P3 Aprendizaje Supervisado

CM: **Decision Tree**



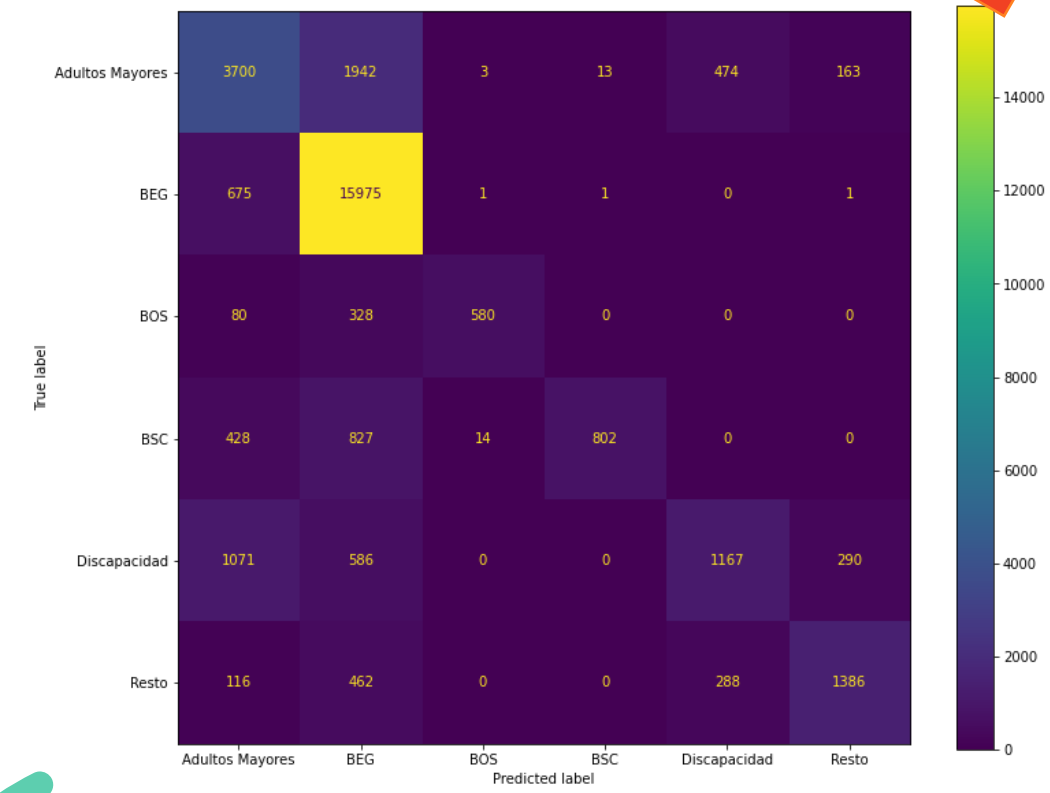
CM: **Random Forest**



P3 Aprendizaje Supervisado

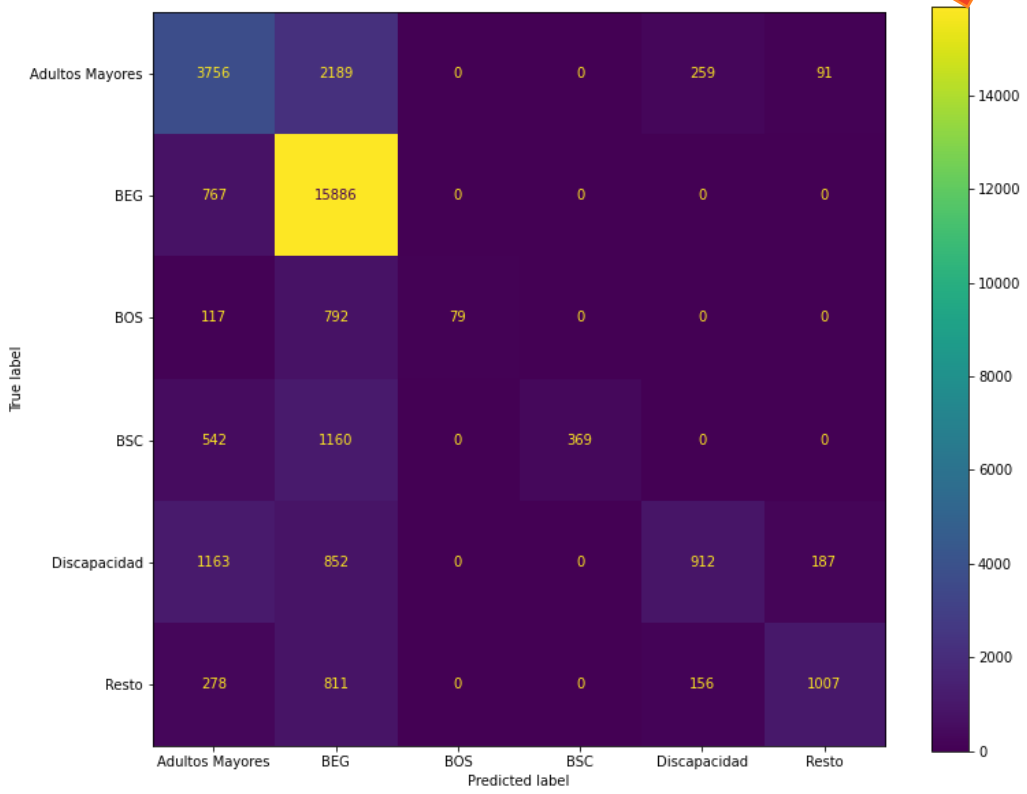
CM: **Decision Tree**

Sin contrato 'Común'



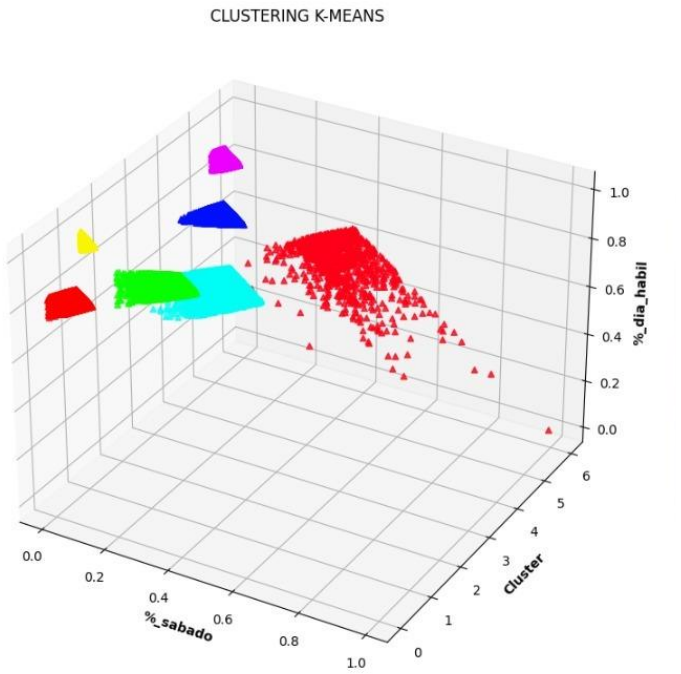
CM: **Random Forest**

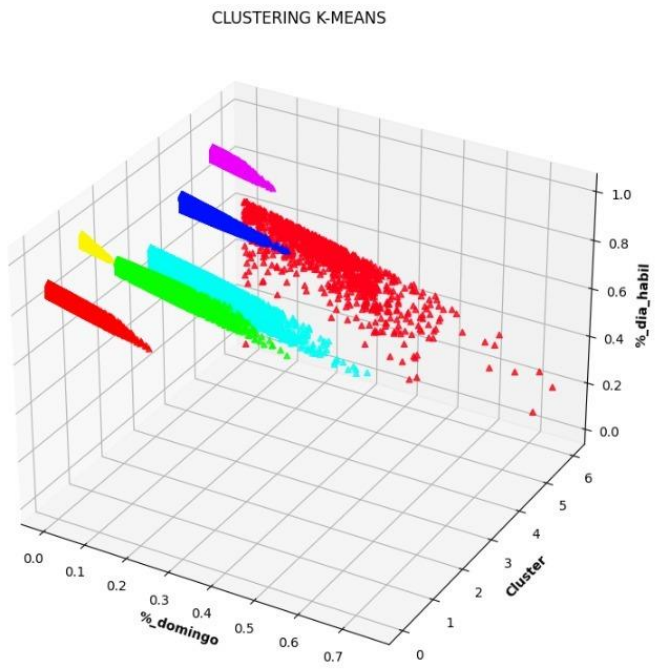
Sin contrato 'Común'

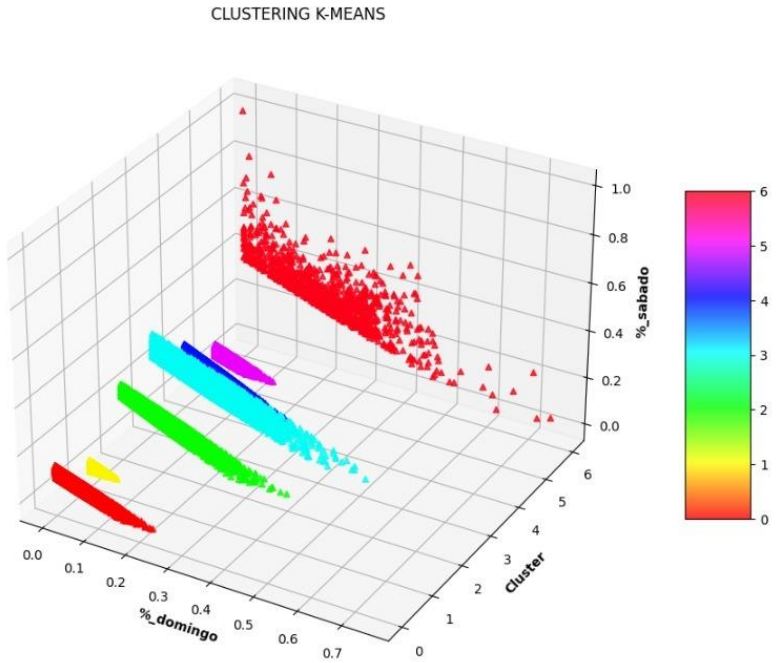


P4 Aprendizaje No Supervisado

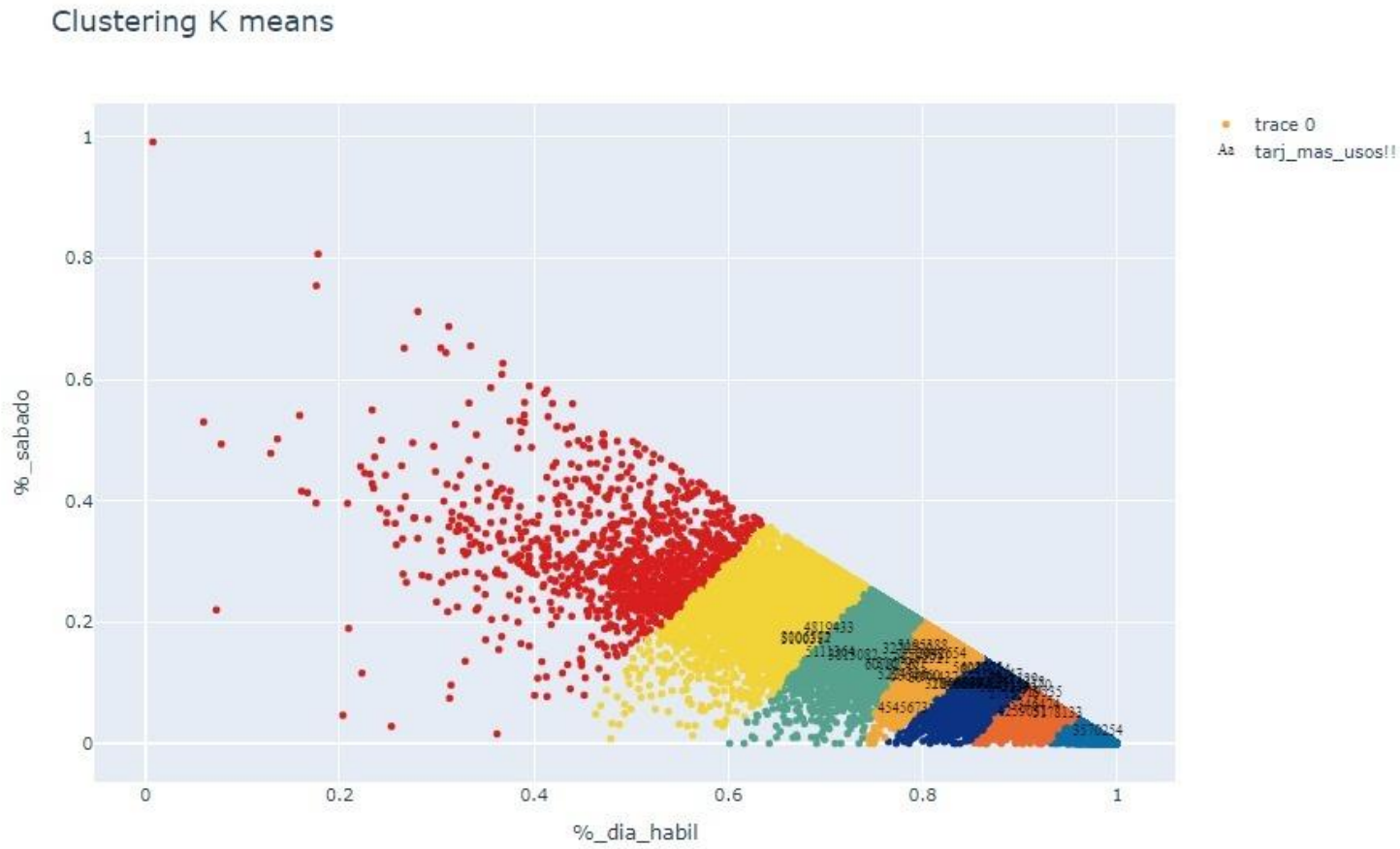
Cluster	Tarjetas	% de tarjetas	% lunes a viernes	% sábados	% domingo	% por grupo	Pasajero
0	23,602	20%	87%	9%	4%	100%	Regular
1	26,460	22%	98%	2%	1%	100%	Regular
2	15,429	13%	74%	15%	10%	100%	Hábil + fin de semana
3	5,604	5%	66%	20%	14%	100%	Hábil + fin de semana
4	22,607	19%	82%	13%	5%	100%	Regular + sábado
5	23,841	20%	93%	5%	2%	100%	Regular
6	1,148	1%	49%	31%	20%	100%	Fin de semana
	118,691	**100%**					







P4 Aprendizaje No Supervisado



Cierre

- **Familiarización** con el dataset
- P1 **Análisis** de la estacionalidad
- P2 **Análisis** del uso de tarjetas
- P3 **Aprendizaje** supervisado
- P4 **Aprendizaje** no supervisado

- Notebooks en repositorio
  - https://github.com/ekocian/mentoria-transporte-urbano-g1