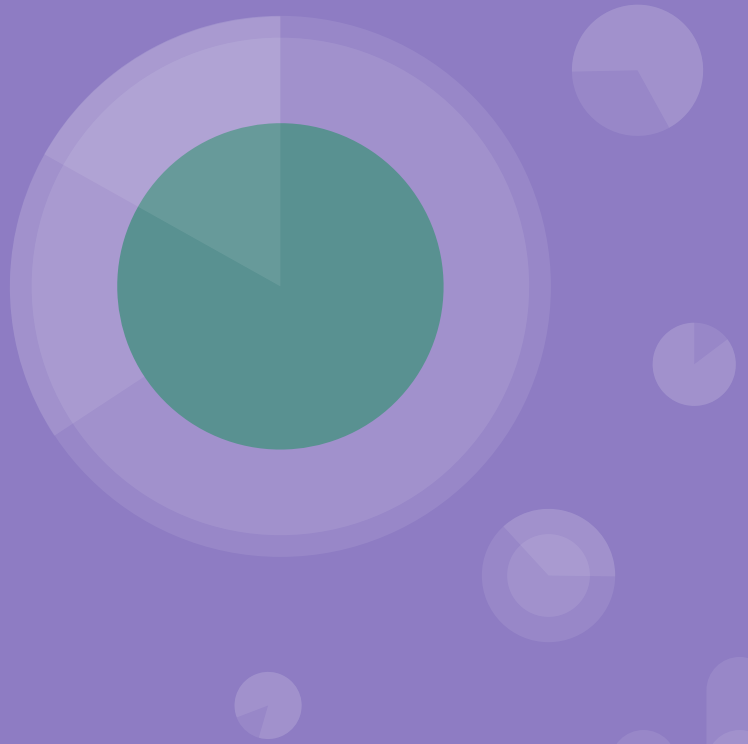
Proyecto Final

Ciencia de Datos



Marco García Duarte

PROYECTO FINAL

- ¿Qué preguntas me motivaron a buscar los datos?

- ¿Qué dataset elegí y dónde lo encontré?

https://data.buenosaires.gob.ar/dataset/

	FECHA	DESDE	HASTA	LINEA	MOLINETE	ESTACION	pax_pagos	pax_pases_pagos	pax_franq	pax_TOTAL
0	1/4/2022	05:15:00	05:30:00	LineaA	LineaA_Congreso_N_Turn01	Congreso	0	0	1	1.0
1	1/4/2022	05:15:00	05:30:00	LineaA	LineaA_Flores_Este_Turn03	Flores	1	0	0	1.0
2	1/4/2022	05:15:00	05:30:00	LineaA	LineaA_Pasco_Turn01	Pasco	0	0	1	1.0
3	1/4/2022	05:15:00	05:30:00	LineaA	LineaA_SanPedrito_Este_Turn04	San Pedrito	2	0	0	2.0
4	1/4/2022	05:15:00	05:30:00	LineaA	LineaA_SanPedrito_Oeste_Turn02	San Pedrito	1	0	0	1.0
...	...	...	...	...	...	...	...	...	...	...
2468473	31/5/2022	23:30:00	23:45:00	LineaH	LineaH_Caseros_Sur_Turn02	Caseros	0	0	1	1.0
2468474	31/5/2022	23:30:00	23:45:00	LineaH	LineaH_Hospitales_Sur_Turn03	Hospitales	1	0	0	1.0
2468475	31/5/2022	23:30:00	23:45:00	LineaH	LineaH_Once_Sur_Turn04	Once	1	0	0	1.0
2468476	31/5/2022	23:30:00	23:45:00	LineaH	LineaH_Cordoba_Turn01	Cordoba	1	0	0	1.0
2468477	31/5/2022	23:30:00	23:45:00	LineaH	LineaH_Cordoba_Turn02	Cordoba	1	0	0	1.0

2468478 rows × 10 columns

PROYECTO FINAL

- Limpieza del dataset
- Dropee las columnas “FECHA”, “DESDE” y “HASTA”, ya que no aportan a la predicción
- Predije la cantidad de pasajeros con un árbol de decisión y una regresión lineal simple

PROYECTO FINAL

```
[24] data_molinetes.head()
```

	LINEA	MOLINETE	ESTACION	pax_TOTAL	DIADELASEMANA	MES	ANO	HORA
0	LineaA	LineaA_Congreso_N_Turn01	Congreso	1.0	4	4	2022	5
1	LineaA	LineaA_Flores_Este_Turn03	Flores	1.0	4	4	2022	5
2	LineaA	LineaA_Pasco_Turn01	Pasco	1.0	4	4	2022	5
3	LineaA	LineaA_SanPedrito_Este_Turn04	San Pedrito	2.0	4	4	2022	5
4	LineaA	LineaA_SanPedrito_Oeste_Turn02	San Pedrito	1.0	4	4	2022	5

```
[54] #Aca iniciamos un modelo de arból de desiciones, importamos el modelo y tomamos la variable a predecir que es pax_TOTAL (la cantidad de pasajeros)

     from sklearn.model_selection import train_test_split
     X = data.drop(columns = ['pax_TOTAL'])
     y = data['pax_TOTAL']
```

```
[55] #Creamos el arbol

     X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=138)  #por default 25% de test
     from sklearn.tree import DecisionTreeRegressor
     arbol = DecisionTreeRegressor()
```

```
[56] arbol.fit(X_train, y_train)

     DecisionTreeRegressor()
```

```
[57] from sklearn.metrics import mean_squared_error
```

```
[58] #Predecimos 'y'
     y_pred = arbol.predict(X_train)
```

```
[59] #Da 0, por lo que hizo over fitting, es decir que esta memorizando los datos que no puede finalizar
     mean_squared_error(y_train, y_pred, squared= False)

     0.0
```

```
[62] y_pred_test = arbol.predict(X_test)
```

```
[63] mean_squared_error(y_test, y_pred_test, squared= False)
```

```
551.6954457490646
```

```
# Aca hacemos un limite de preguntas para que el modelo dé valores razonables de pasajeros
arbol_2 = DecisionTreeRegressor(max_depth= 15)
arbol_2.fit(X_train, y_train)
```

```
DecisionTreeRegressor(max_depth=15)
```

```
[91] y_pred = arbol_2.predict(X_train)
     mean_squared_error(y_train, y_pred, squared= False)
```

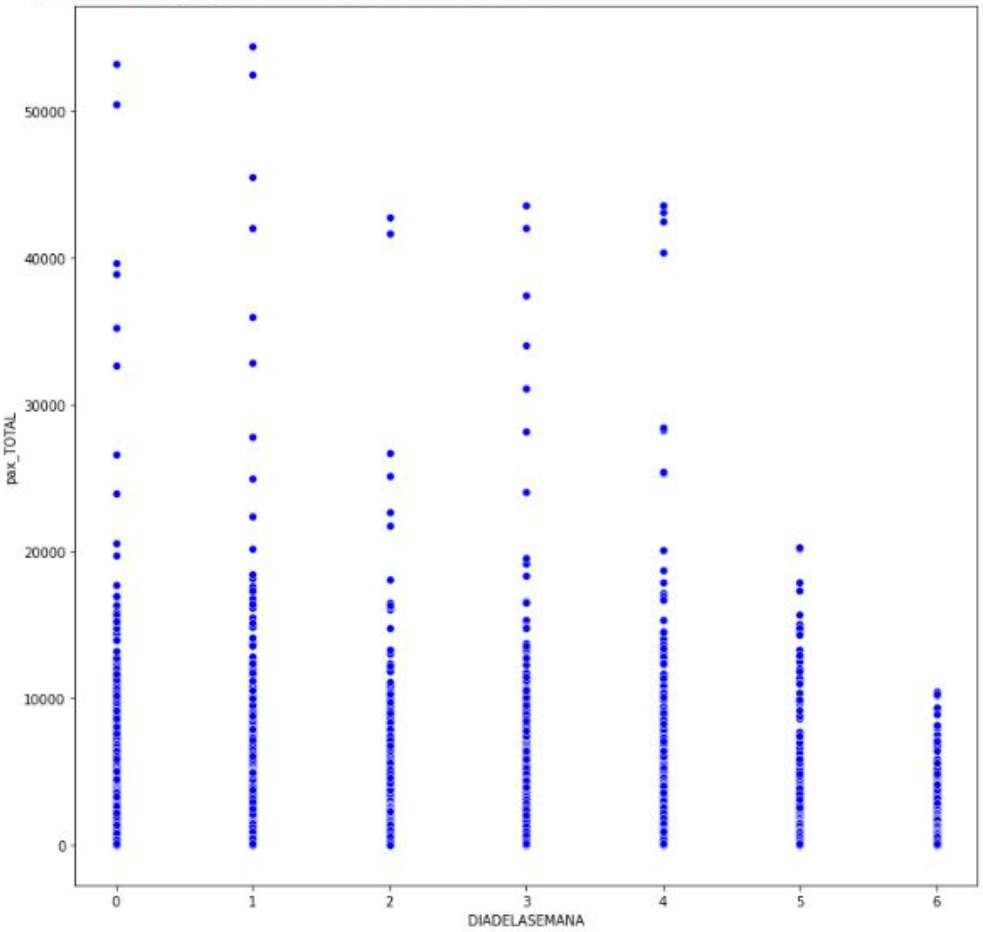
```
903.9532806536458
```

```
[92] y_pred_test = arbol_2.predict(X_test)
```

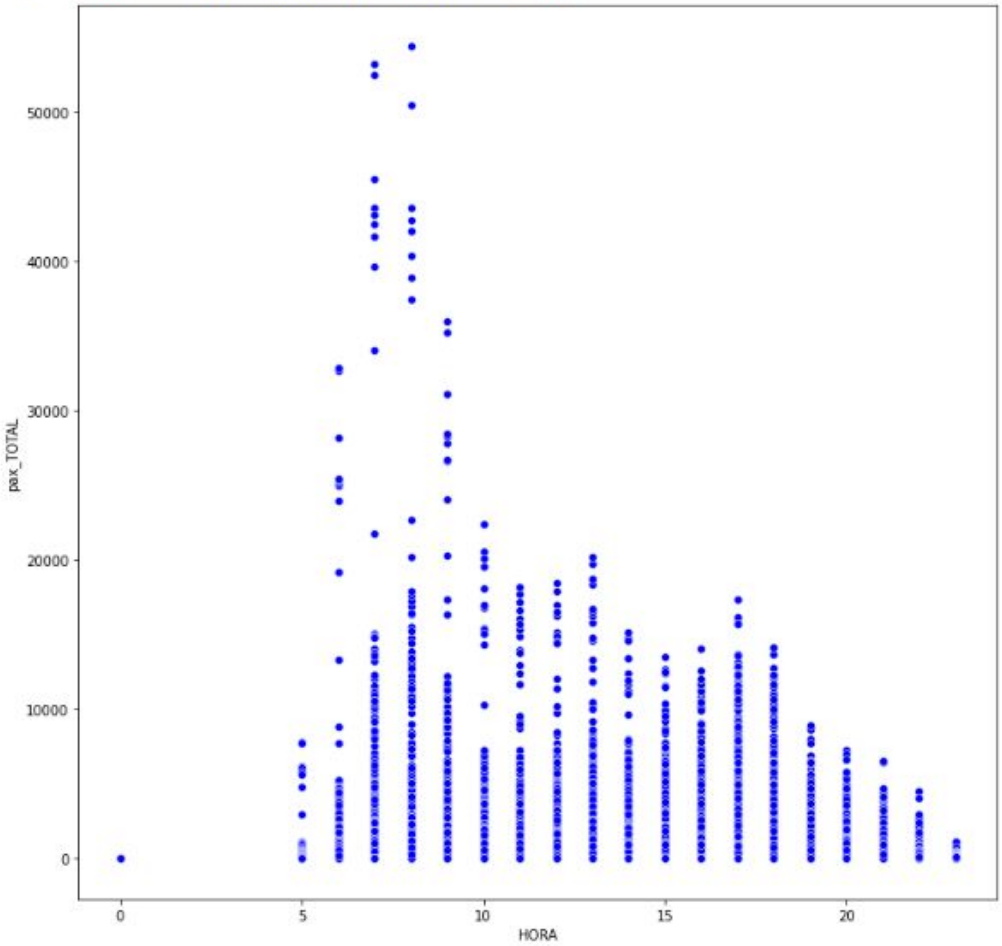
```
[93] mean_squared_error(y_test, y_pred_test, squared= False)
```

```
1048.3661050425892
```

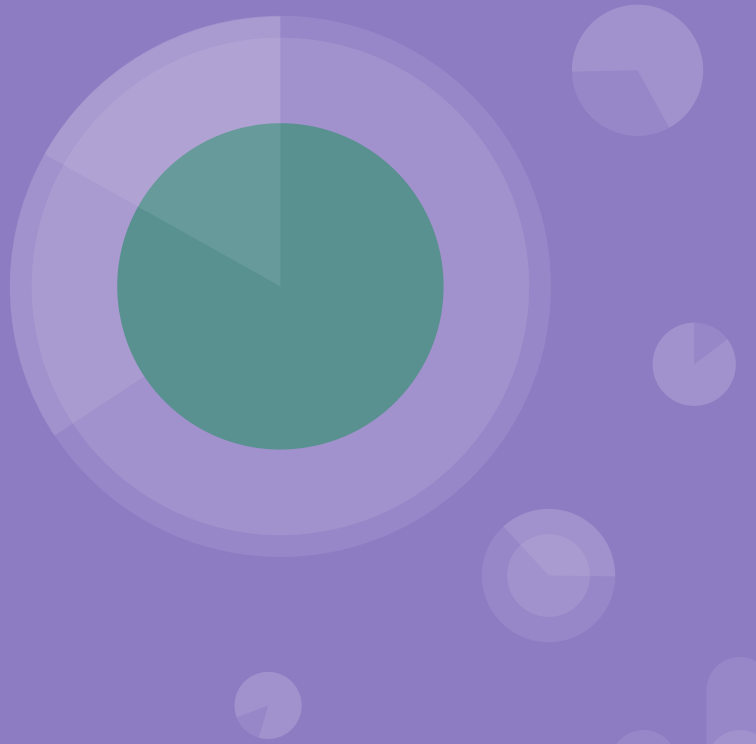
<matplotlib.axes._subplots.AxesSubplot at 0x7f44124230a0>



<matplotlib.axes._subplots.AxesSubplot at 0x7f4412423f70>



CONCLUSIONES GENERALES



Es rentable montar un negocio, ya sea de comida, indumentarias, etc, debido a la cantidad de pasajeros que circulan