Didi Business Intelligence Challenge

The following document contains a detailed explanation of the procedure followed to get the answers to the questions 1 to 4 of the Technical Skills Challenge.

Procedure Steps:

1. Data Exploration of CSV files using Excel
2. Noticed an error on visit_date column in the 'restaurants_visitors' file.

1	id	visit_date	visit_datetime	reserve_datetime	reserve_
1567	24b9b2a0	7/26/2016	26/07/2016 17:00	25/07/2016 21:00	2
1568	aed3a8b4	9/12/2016	12/9/2016 15:00	12/8/2016 16:00	2
1569	965b2e0c	#VALUE!	26/10/2016 20:00	25/10/2016 17:00	5
1570	42c9aa6d	#VALUE!	27/10/2016 11:00	26/10/2016 13:00	5
1571	45326ebb	#VALUE!	27/10/2016 11:00	26/10/2016 21:00	1
1572	45326ebb	#VALUE!	27/10/2016 11:00	26/10/2016 21:00	1
1573	0a74a540	#VALUE!	27/10/2016 17:00	27/10/2016 14:00	2
1574	0a74a540	#VALUE!	27/10/2016 17:00	27/10/2016 14:00	2
1575	24b9b2a0	#VALUE!	27/10/2016 17:00	22/10/2016 18:00	6

3. Filled the cells that had #VALUE! text with the date from the visit_datetime column by coping the entire column and changing the format to 'YYYY-MM-DD'.
4. Changed the format of the columns, visit_datetime and reserve_datetime, from 'DD-MM-YY HH:MM:SS' to 'YYYY-MM-DD HH:MM:SS' to match them with SQL datetime variable format.
5. Creation of tables and dependencies on SQL Server Management Studio.

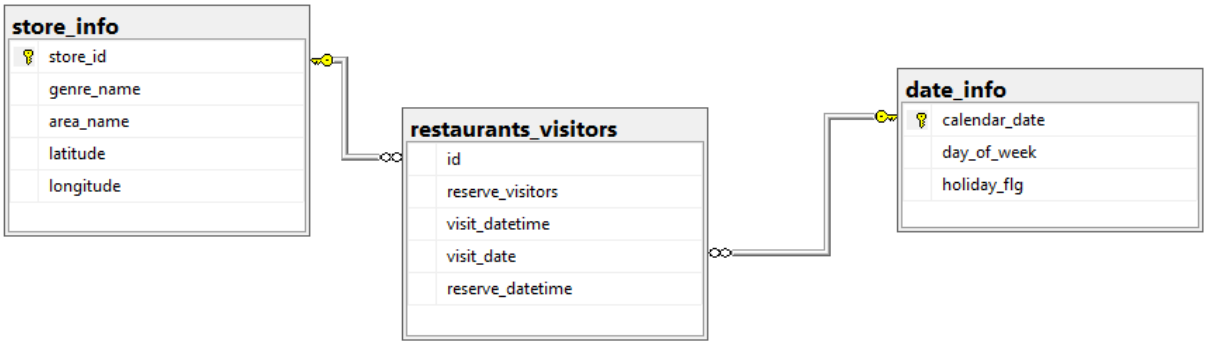
```
-- Creación de la Base de Datos
CREATE DATABASE DIDI
USE DIDI

-- Tabla date_info:
CREATE TABLE date_info (
calendar_date date,
day_of_week varchar(10), -- El valor mas grande es de 9 caracteres y promedio de 7
holiday_flg int,
PRIMARY KEY(calendar_date))
```

```sql
-- Tabla store_info:
CREATE TABLE store_info (
store_id char(16), -- Todos los valores son de 16 caracteres
genre_name varchar(50), -- El valor mas grande es de 28 caracteres y promedio de 28
area_name nvarchar(50), -- El valor mas grande es de 49 caracteres y promedio de 32
latitude float,
longitude float,
PRIMARY KEY (store_id))

-- Tabla restaurants_visitors:
CREATE TABLE restaurants_visitors (
id char(16), -- Todos los valores son de 16 caracteres
reserve_visitors int,
visit_datetime datetime,
visit_date date,
reserve_datetime datetime,
FOREIGN KEY (id) REFERENCES store_info(store_id),
FOREIGN KEY (visit_date) REFERENCES date_info(calendar_date)
);
```

6. Creation of ER diagram to visualize the tables connections.



7. Data Import into the created tables.
8. During Data Exploration, discovered that the first record is incorrect because it's the only record where the reservation time occurs after the visit time, which isn't logical, therefore that row was deleted.

```sql
-- !! REGISTRO INCORRECTO (Fecha de visita es menor que la fecha de reserva)
SELECT *
FROM restaurants_visitors
WHERE visit_datetime < reserve_datetime
-- Borrar registro equivocado
DELETE FROM restaurants_visitors
WHERE visit_datetime < reserve_datetime
```

9. Wrote queries for more data exploration and answered the SQL questions of the Didi Business Intelligence Challenge.
10. Import data into Visual Studio Code.
11. With pandas, sklearn, and statsmodels read the csv file and created a data frame.
12. Grouped reserve_visitors column by Months, to train a model and forecast the next 6 months of the sum of visitors of all restaurants.
13. Discovered there are no records for August 2016 (2016-08) and atypical data for September 2016, and May 2017. Also, there is a huge increment on visitors from November 2016 (2016-11) to April 2017 (2017-04)

	visit_datetime	reserve_visitors
0	2016-01	906
1	2016-02	868
2	2016-03	1307
3	2016-04	1340
4	2016-05	833
5	2016-06	1205
6	2016-07	1280
7	2016-09	2
8	2016-10	901
9	2016-11	4809

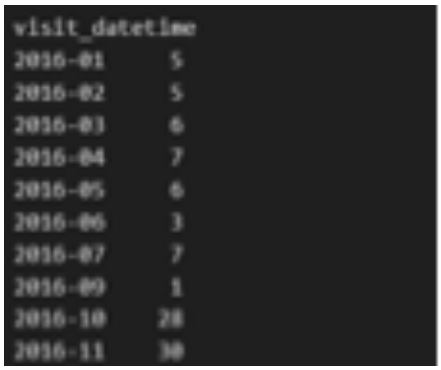
5	2016-06	1205
6	2016-07	1280
7	2016-09	2
8	2016-10	901
9	2016-11	4809
10	2016-12	9785
11	2017-01	5998
12	2017-02	6077
13	2017-03	8152
14	2017-04	5228
15	2017-05	452

14. Checked the groups' sizes to see how many rows of data were per month. Discovered anomalies on the months that matched the previous ones identified.

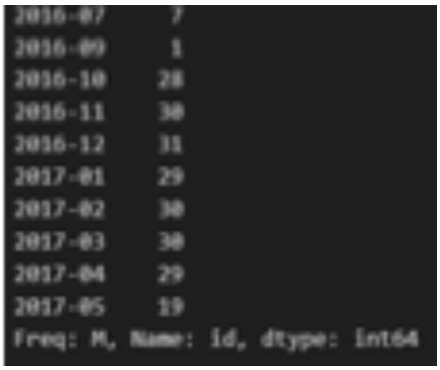
```
visit_datetime
2016-01     203
2016-02     202
2016-03     269
2016-04     277
2016-05     169
2016-06     182
2016-07     264
2016-09       1
2016-10     239
```

```
2016-09       1
2016-10     239
2016-11    1323
2016-12    2324
2017-01    1551
2017-02    1707
2017-03    2024
2017-04    1335
2017-05      92
Freq: M, dtype: int64
```

15. Checked how many restaurants were per month and discovered that starting from October 2016 (2016-10), the number of restaurants incremented.



```
visit_datetime
2016-01     5
2016-02     5
2016-03     6
2016-04     7
2016-05     6
2016-06     3
2016-07     7
2016-09     1
2016-10    28
2016-11    30
```



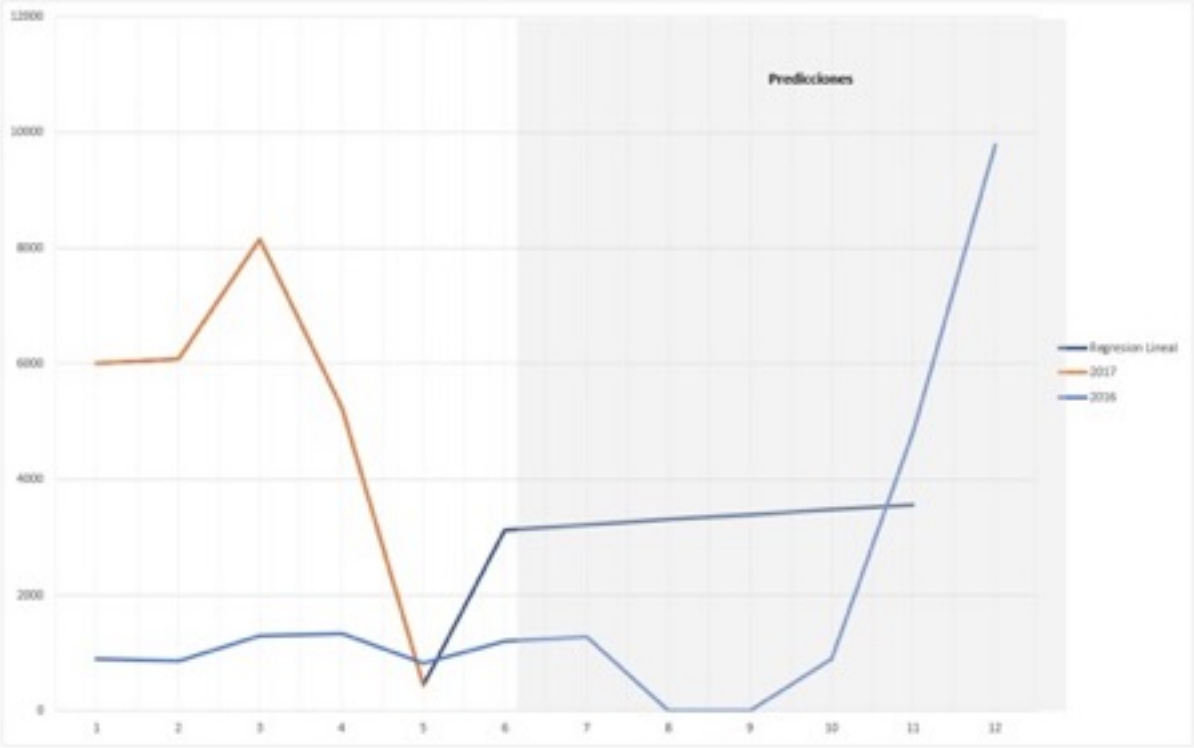
```
2016-07     7
2016-09     1
2016-10    28
2016-11    30
2016-12    31
2017-01    29
2017-02    30
2017-03    30
2017-04    29
2017-05    19
Freq: M, Name: id, dtype: int64
```

16. Forecast the next 6 months using Linear Regression Model.

	visit_datetime	Predicted Visitors
0	2017-06-30	3130.711976
1	2017-07-31	3216.929396
2	2017-08-31	3303.146816
3	2017-09-30	3389.364236
4	2017-10-31	3475.581656
5	2017-11-30	3561.799076

17. Input these data points to excel and created the following graphs.

	Regresion Lineal									
	Año	Mes	SUM Visitantes	%	Año	Mes	SUM Visitantes	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	452	-91%	-381	54
	2016	6	1205	45%	2017	6	3131	593%	1926	548
	2016	7	1280	6%	2017	7	3217	3%	1937	3
	2016	8	0	-100%	2017	8	3303	3%	3303	103
	2016	9	2	0%	2017	9	3389	3%	3387	3
	2016	10	901	44950%	2017	10	3476	3%	2575	44947
	2016	11	4809	434%	2017	11	3562	2%	-1247	431
	2016	12	9785	103%	2017	12				
Promedio			1936	4132%			4180	43%		9121

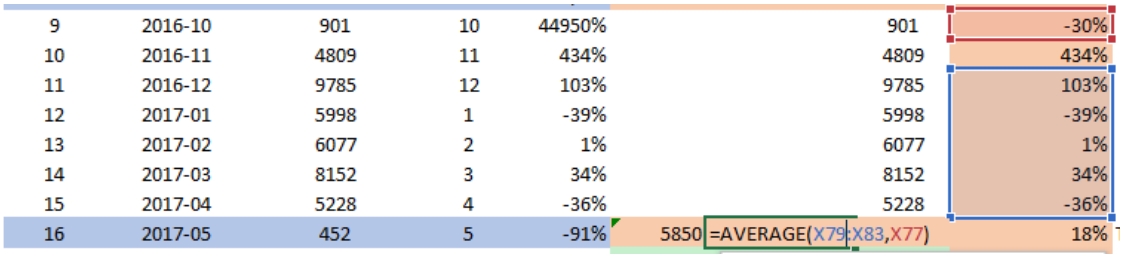


18. Compared the differences between sum of visitors and percentual change per month, of years 2016 vs 2017, and analyzed the pattern that the graph showed to determine that a Linear Regression Model wasn't accurate enough to forecast the next 6 months of total visitors.
19. Transformed atypical data using Feature Engineering for the following months:
    - August 2016 (2016-08)
    - September 2016 (2016-09)
    - May 2017 (2017-05)

20. The criteria used for the Feature Engineering was to get the average of total visitors of the 7 previous months, and the average of the percentage growth for the 6 previous months of the record to change. Then multiply the average total visitors times the average of percentage growth to get the new value.

Ajuste Valores Atípicos (Promedios y % Crecimiento)								
	visit_datetime	reserve_visitors	month	%	g 7 anterior Avg	% 6 anteriores	New Sum	New %
0	2016-01	906	1				906	
1	2016-02	868	2	-4%			868	-4%
2	2016-03	1307	3	51%			1307	51%
3	2016-04	1340	4	3%			1340	3%
4	2016-05	833	5	-38%			833	-38%
5	2016-06	1205	6	45%			1205	45%
6	2016-07	1280	7	6%			1280	6%
7	2016-08	0	8	-100%	1106	10%	1220	-5%
8	2016-09	2	9	#DIV/0!	1134	13%	1279	5%
9	2016-10	901	10	44950%			901	-30%
10	2016-11	4809	11	434%			4809	434%
11	2016-12	9785	12	103%			9785	103%
12	2017-01	5998	1	-39%			5998	-39%
13	2017-02	6077	2	1%			6077	1%
14	2017-03	8152	3	34%			8152	34%
15	2017-04	5228	4	-36%			5228	-36%
16	2017-05	452	5	-91%	5850	6%	6190	18%

21. Observed that the anomalies caused an unusual percentage growth for November 2016 (2016-11), so that datapoint was excluded from the average, instead the previous one, calculated by the new transformed data, was used.



22. Compared the new calculated value for May 2017 (2017-05) with its previous year performance. This new value had an increased number of total visitors and a percentage growth of 18% which wasn't coherent with 2016 performance where the total visitors had decreased from its previous month and its percentage growth was of -38%.
23. Recalculate this value using an alternate procedure.
24. Averaged the number of records for the last 7 months and divided the total visitors by the number of records per month to get the relation between visitors per record. Averaged the relation and multiply it to get an estimated value of total visitors.
25. Averaged the percentage growth for the records of the first 4 months of both years.

26. Compared them and decided to use the average for year 2017.
27. Multiplied the average records by the average percentage growth, to adjust the value considering growth.
28. Multiplied this value by the average relation of visitors by record of the last 7 months.
29. This new estimation of Total Visitors presented a percentage growth of -0.1% which made more sense than an increment of 18% of growth.

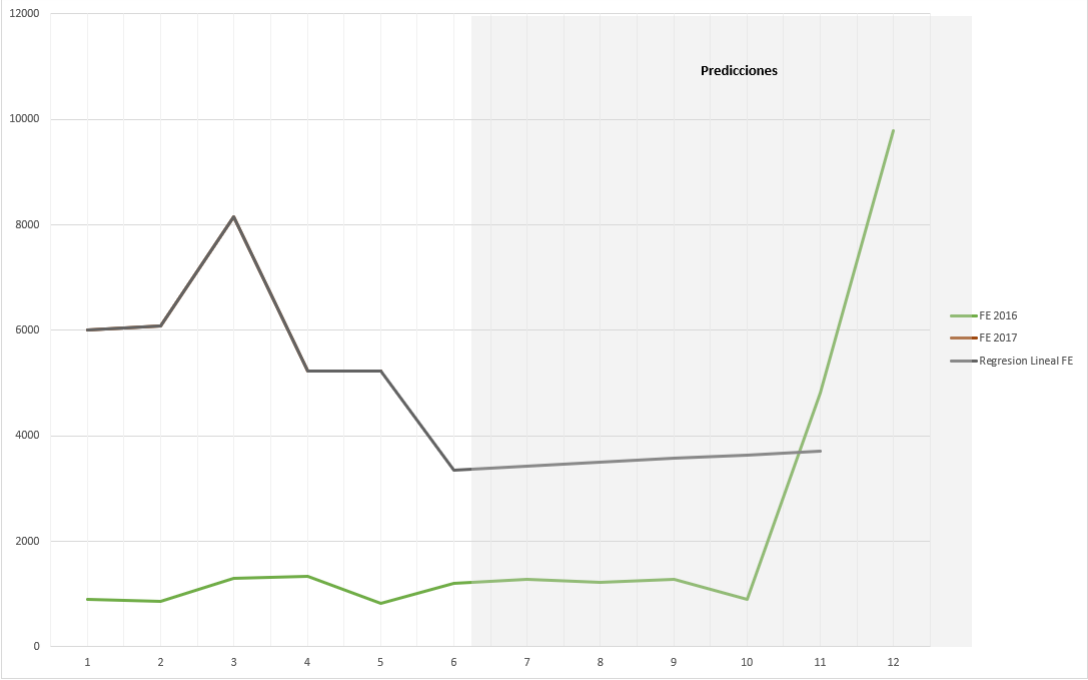
15	2017-04	3228	4	-36%			3228	-36%
16	2017-05	452	5	-91%	5850	6%	6190	18%
					5225		5225	0%

30. Forecast the next 6 months using Linear Regression Model with transformed data.

	visit_datetime	Predicted Visitors
0	2017-06-30	3355.940647
1	2017-07-31	3426.828537
2	2017-08-31	3497.716427
3	2017-09-30	3568.604317
4	2017-10-31	3639.492206
5	2017-11-30	3710.380096

31. Input these data points to excel and created the following graphs.

	Regresion Lineal FE									
	Año	Mes	SUM Visitant	%	Año	Mes	SUM Visitant	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	5225	0%	4392	38
	2016	6	1205	45%	2017	6	3356	-36%	2151	80
	2016	7	1280	6%	2017	7	3427	2%	2147	4
	2016	8	1220	-5%	2017	8	3498	2%	2278	7
	2016	9	1279	0%	2017	9	3568	2%	2289	2
	2016	10	901	-30%	2017	10	3639	2%	2738	32
	2016	11	4809	434%	2017	11	3710	2%	-1099	432
	2016	12	9785	103%	2017	12				
Promedio			2144	51%			4716	-6%		25



32. This forecast had a more stable prediction of total visitors than the previous Linear Regression Model.
33. Checked another algorithm to compare accuracy.
34. Decided to try an ARIMA model, because of the increment of restaurants recorded that started from October 2016 (2016-10).

35. Created 2 forecasts using ARIMA, one with the original data and one with the transformed data.

forecast for original data
2254.169451
3875.105115
2122.640994
2691.491263
5077.760626
4295.300489

forecast for original data

forecast for transformed data
7178.179648
5495.715259
5465.951435
6844.953476
5663.457952
5559.774187

forecast for transformed data

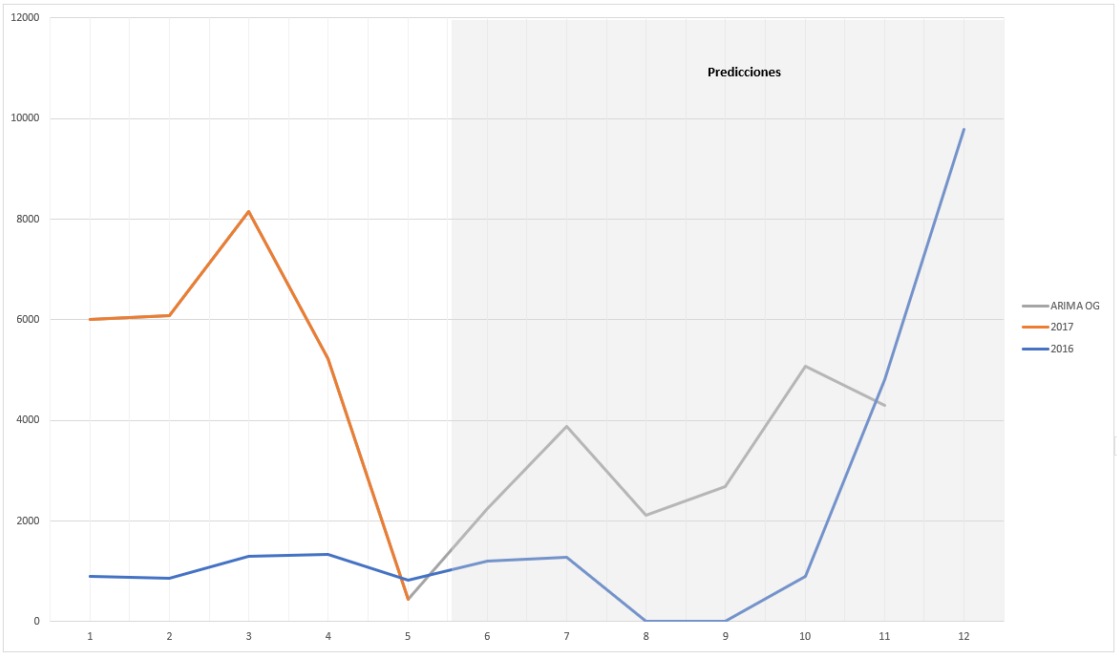
36. Input these data points to excel and created the following graphs.

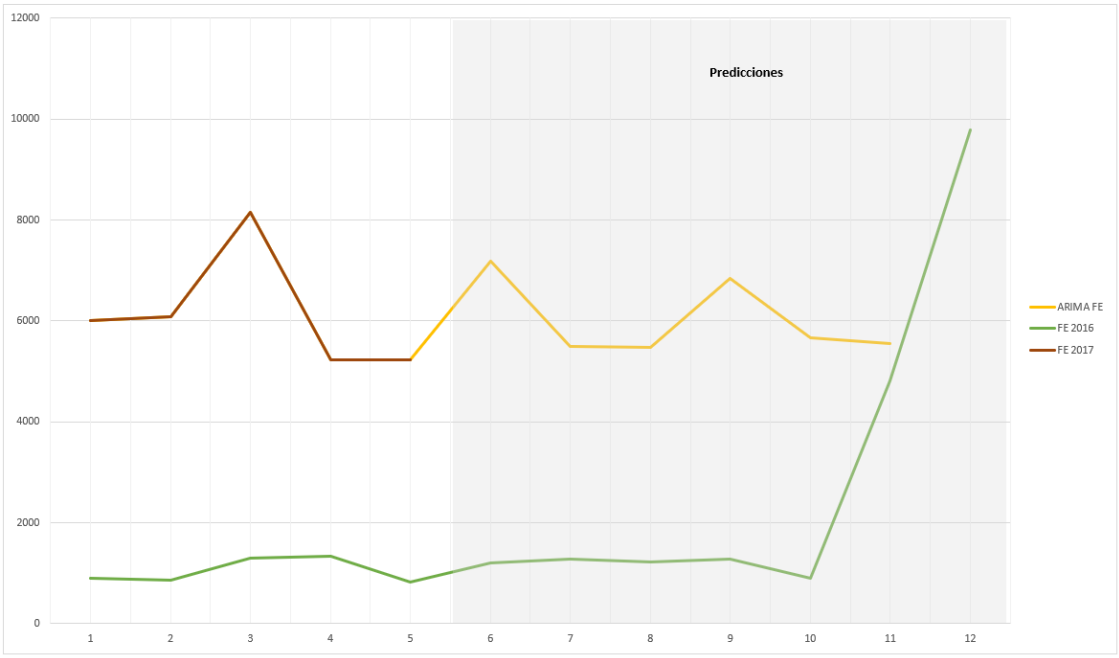
Crecimiento Mensual (ARIMA)

	Año	Mes	JM Visitant	%	Año	Mes	JM Visitant	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	452	-91%	-381	54
	2016	6	1205	45%	2017	6	2254	399%	1049	354
	2016	7	1280	6%	2017	7	3875	72%	2595	66
	2016	8	0	-100%	2017	8	2123	-45%	2123	55
	2016	9	2	0%	2017	9	2691	27%	2689	27
	2016	10	901	44950%	2017	10	5078	89%	4177	44861
	2016	11	4809	434%	2017	11	4295	-15%	-514	449
	2016	12	9785	103%	2017	12				
Promedio			1936	4132%			4202	36%		9073

Crecimiento Mensual (ARIMA FE)

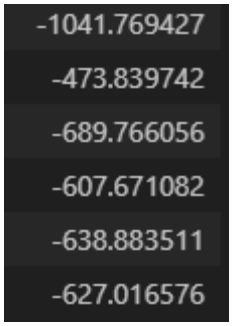
	Año	Mes	JM Visitant	%	Año	Mes	JM Visitant	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	5225	0%	4392	38
	2016	6	1205	45%	2017	6	7178	37%	5973	7
	2016	7	1280	6%	2017	7	5495	-23%	4215	30
	2016	8	1220	-5%	2017	8	5466	-1%	4246	4
	2016	9	1279	0%	2017	9	6845	25%	5566	25
	2016	10	901	-30%	2017	10	5663	-17%	4762	12
	2016	11	4809	434%	2017	11	5559	-2%	750	436
	2016	12	9785	103%	2017	12				
Promedio			2144	51%			6081	-2%		16



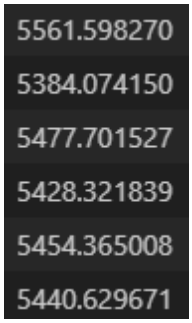


37. The graphs showed that these models adjust better than the Linear Regression Models.
38. The prediction of the model trained with original data shows a similar pattern on the months where anomalies present in 2016.
39. The prediction of the model trained with FE data had a spikes pattern that didn't showed in 2016, but it maintains a more stable total of visitors throughout 2017.
40. Checked if by changing the hyperparameters the ARIMA model could be more accurate.

41. Did a Time Series Cross Validation using TimeSeriesSplit (TSS) to evaluate the model's RMSE using different hyperparameters.
42. Comparing the hyperparameters (5,1,0) and (1,1,1), the one that had the lowest Root Mean Squared Error (RMSE), was (1,1,1). *The previous 2 predictions used (5,1,0)***
43. Forecast the next 6 months using ARIMA Model with the new hyperparameters (1,1,1) with both the original and transformed data.



forecast for original data



forecast for transformed data

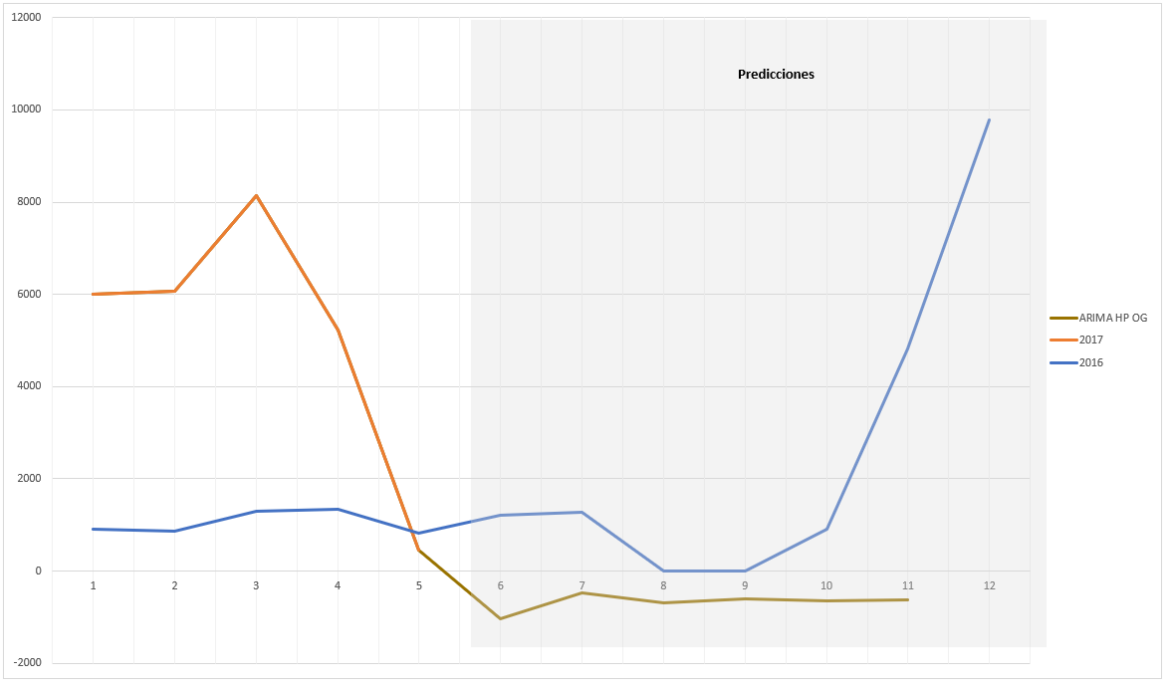
44. Input these data points to excel and created the following graphs.

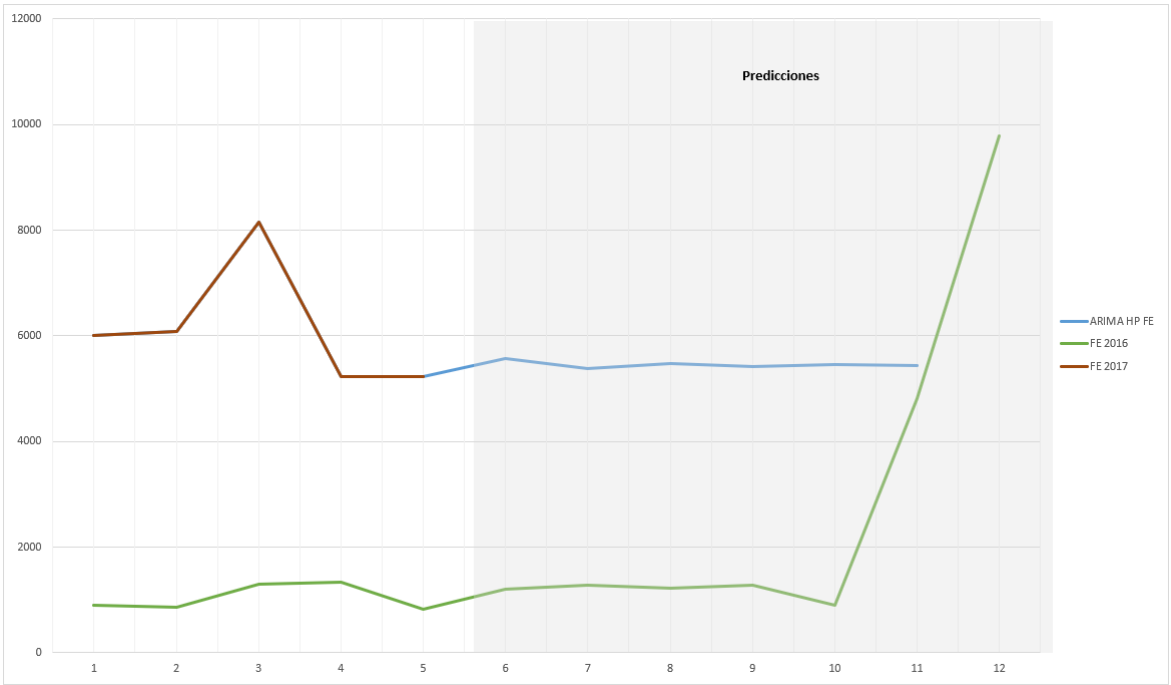
Crecimiento Mensual (ARIMA, Hiperparametros)

	Año	Mes	UM Visitant	%	Año	Mes	UM Visitant	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	452	-91%	-381	54
	2016	6	1205	45%	2017	6	-1042	-331%	-2247	375
	2016	7	1280	6%	2017	7	-474	-55%	-1754	61
	2016	8	0	-100%	2017	8	-690	46%	-690	146
	2016	9	2	0%	2017	9	-608	-12%	-610	12
	2016	10	901	44950%	2017	10	-639	5%	-1540	44945
	2016	11	4809	434%	2017	11	-627	-2%	-5436	436
	2016	12	9785	103%	2017	12				
Promedio			1936	4132%			1984	-44%		9108

Crecimiento Mensual (ARIMA, Hiperparametros y FE)

	Año	Mes	UM Visitant	%	Año	Mes	UM Visitant	%	Diferencias SUM	Diferencias %
	2016	1	906		2017	1	5998	-39%	5092	39
	2016	2	868	-4%	2017	2	6077	1%	5209	6
	2016	3	1307	51%	2017	3	8152	34%	6845	16
	2016	4	1340	3%	2017	4	5228	-36%	3888	38
	2016	5	833	-38%	2017	5	5225	0%	4392	38
	2016	6	1205	45%	2017	6	5562	6%	4357	38
	2016	7	1280	6%	2017	7	5384	-3%	4104	9
	2016	8	1220	-5%	2017	8	5478	2%	4258	6
	2016	9	1279	0%	2017	9	5428	-1%	4149	1
	2016	10	901	-30%	2017	10	5454	0%	4553	30
	2016	11	4809	434%	2017	11	5441	0%	632	434
	2016	12	9785	103%	2017	12				
Promedio			2144	51%			5766	-3%		17





45. The prediction of the model trained with original data returned negative values.
46. The prediction of the model trained with FE data presented a similar pattern to the year 2016.

47. To select the best modeling the Accuracy Metric was defined by the Average Percentual Change for the last 5 months between 2016 data and the 2017 predicted data. This is because it is the comparison between the predicted months and original data from its previous year, excluding November because its unusual percentage growth.

2016	5	833	-38%	2017	5	5223	0%	4392	38
2016	6	1205	45%	2017	6	7178	37%	5973	7
2016	7	1280	6%	2017	7	5495	-23%	4215	30
2016	8	1220	-5%	2017	8	5466	-1%	4246	4
2016	9	1279	0%	2017	9	6845	25%	5566	25
2016	10	901	-30%	2017	10	5663	-17%	4762	12
2016	11	4809	434%	2017	11	5559	-2%	750	436
2016	12	9785	103%	2017	12				
edio		2144	51%			6081	-2%		=AVERAGE(BG55:BG59)

48. The model with the hyperparameters (5,1,0) had a lower average percentual change (16) than the model with hyperparameters (1,1,1) that had (17).
49. The model selected was number 4. (ARIMA with FE data, with (5,1,0) hyperparameters)

Diferencias %
39
6
16
38
38
7
30
4
25
12
436
16

Model 4

Diferencias %
39
6
16
38
38
38
9
6
1
30
434
17

Model 6