Tableau Data Visualization CA

Cesar Marrades Cortés

Contents

1. Summary 3

2. Background 3

3. Dataset – Record Sample 4

4. Seven stages 4

4.1. Acquiring Stage 4

4.2. Parsing Stage 4

CarparkCapacity 5

Removing records surpassing max capacity 5

GeoLocating Datapoints 6

4.3. Filtering Stage 6

4.4. Mining Stage – Removing Extra rows 6

4.5. Represent Stage 7

Main Dashboard 7

Yearly Occupancy Ratios 8

Occupancy Trends Dashboard 8

Monthly Breakdown By area 9

Monthly Breakdown by Parking 10

4.6. Refining Stage 10

AVG occupancies over 100% 10

Extra tooltips 11

Occupancy Ratio 11

Ratio as Percentage 11

More calculated fields 11

Tooltips must show valuable information 12

A graph missing something 12

Swapping rows per columns 12

Accurate info on Titles 13

Contrast 13

4.7. Interact Stage 14

5. Problems and Solutions 14

Different bubble sorts for same data 14

Bubbles vs Treemaps 16

Comparing data together 17

Data calculations not matching 17

7. Conclusion 18

1. Summary

As an employee of a company that has big amounts of data sitting and waiting to be mined, with the permission of my manager I have decided to utilize it to give it some value. The information presented here is available to the public although it´s an internal process that has allowed me to access to a bigger set of this data.

My initial approach was to work with **two datasets**. At some point I ran into loads of problems due to the volume of the data I was trying to mine for this specific dataset. The 40Gbs of the **Sql** database made a local copy an impossible task. When I managed to import a set of the data, Tableau didn´t find easy to work with it, as one of the tables was containing more than 30 million records.

Eventually the second dataset became the "primary" for this project, and the first one was discarded after hours of work lost, so additional time was spent trying to obtain meaningful information from the visualizations of the second dataset.

The information contained on this CA is publicly updated each 5 minutes on http://www.dublincity.ie/dublintraffic/cpdata.xml, containing free parking spaces data points. The main objective is try to discover trends peak hours, and somehow understand any correlation between all the data presented. As a matter of fact this CA will be used for business perspectives leading to further investigations and strategic decisions.

This document details the process followed in order to represent the visualizations submitted together along this file.

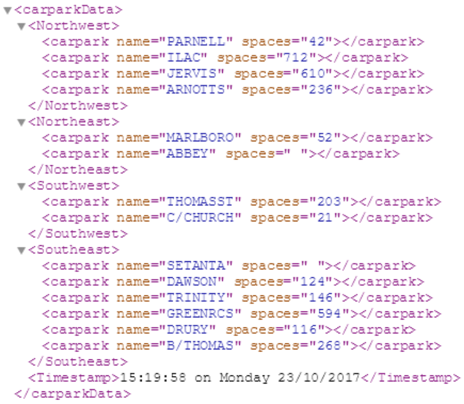
2. Background

Somehow this has been an opportunity to finish an end to end process that started on 2016. By then, I wrote a small process that collects, parses and stores this data every 10 minutes (still live). This information has been since then sitting there with no usage at all. What are the peak times on public parkings? What are the off peak times? Are there any trends? This information may not be useful for most of the people, but it turns to be valuable for our company.

I also would like to focus my final project on something related to this area e.g. trends, predictions based on locations or parking data. So I thought this was a good opportunity to make this data useful to the business at the same time that I was gaining skills on this specific field.

Nobody in my company has ever accessed to this data yet, therefore it is the first time that specific dataset is mined. I am not aware of other parties accessing, storing and mining data though.

3. Dataset – Record Sample



```
▼<carparkData>
  ▼<Northwest>
     <carpark name="PARNELL" spaces="42"></carpark>
     <carpark name="ILAC" spaces="712"></carpark>
     <carpark name="JERVIS" spaces="610"></carpark>
     <carpark name="ARNOTTS" spaces="236"></carpark>
   </Northwest>
  ▼<Northeast>
     <carpark name="MARLBORO" spaces="52"></carpark>
     <carpark name="ABBEY" spaces=" "></carpark>
   </Northeast>
  ▼<Southwest>
     <carpark name="THOMASST" spaces="203"></carpark>
     <carpark name="C/CHURCH" spaces="21"></carpark>
   </Southwest>
  ▼<Southeast>
     <carpark name="SETANTA" spaces=" "></carpark>
     <carpark name="DAWSON" spaces="124"></carpark>
     <carpark name="TRINITY" spaces="146"></carpark>
     <carpark name="GREENRCS" spaces="594"></carpark>
     <carpark name="DRURY" spaces="116"></carpark>
     <carpark name="B/THOMAS" spaces="268"></carpark>
   </Southeast>
   <Timestamp>15:19:58 on Monday 23/10/2017</Timestamp>
 </carparkData>
```

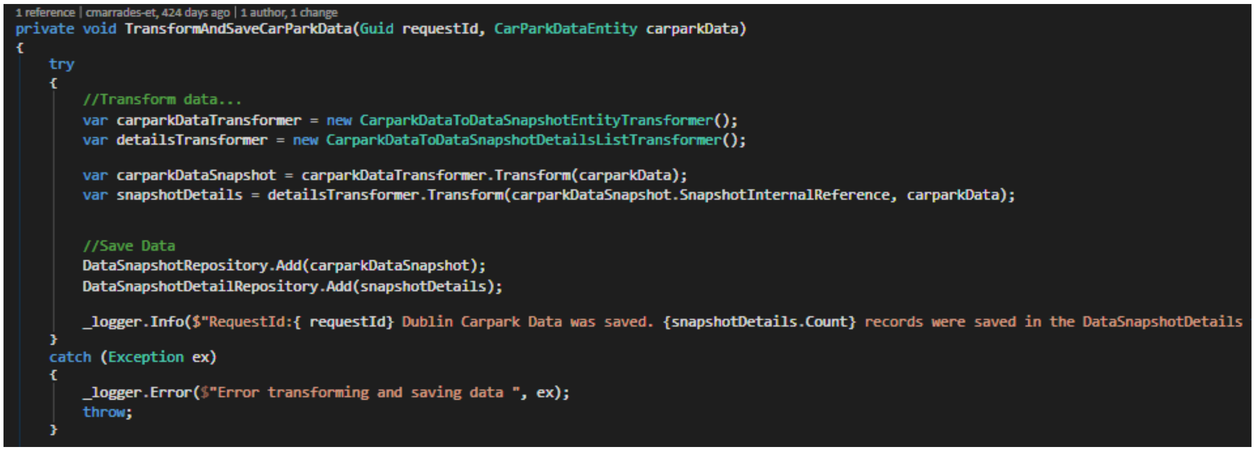
Although Tableau project has been made connecting to a local Sql database, csv files have been generated through Tableau and submitted with this document.

An xml sample has been attached to the whole solution.

4. Seven stages

4.1. Acquiring Stage

As mentioned before, Dublin Car Park data has been collected and stored for the last year every 10 minutes. I personally wrote the application that consumes this information and stores it on a database.



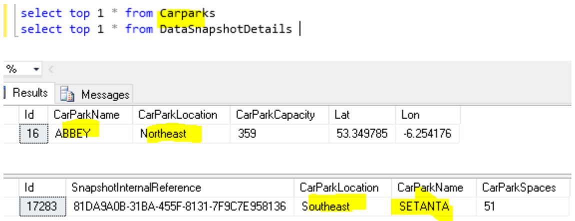
```
1 reference | cmarrades-et, 424 days ago | 1 author, 1 change
private void TransformAndSaveCarParkData(Guid requestId, CarParkDataEntity carparkData)
{
    try
    {
        //Transform data...
        var carparkDataTransformer = new CarparkDataToDataSnapshotEntityTransformer();
        var detailsTransformer = new CarparkDataToDataSnapshotDetailsListTransformer();

        var carparkDataSnapshot = carparkDataTransformer.Transform(carparkData);
        var snapshotDetails = detailsTransformer.Transform(carparkDataSnapshot.SnapshotInternalReference, carparkData);


        //Save Data
        DataSnapshotRepository.Add(carparkDataSnapshot);
        DataSnapshotDetailRepository.Add(snapshotDetails);

        _logger.Info($"RequestId:{ requestId} Dublin Carpark Data was saved. {snapshotDetails.Count} records were saved in the DataSnapshotDetails
    }
    catch (Exception ex)
    {
        _logger.Error($"Error transforming and saving data ", ex);
        throw;
    }
```

On this case I imported locally the tables that contained all the information.

4.2. Parsing Stage

Most part of the parsing job is done by the application that transforms the XML and stores it on the Database.

I found the structure of the tables is not optimal. There is no ParkingId or LocationId on the "SnapshotDetails" table. If this was our final model I would invest some more time on normalizing this structure, as it is not efficient to join both tables by string columns.



```sql
select top 1 * from Carparks
select top 1 * from DataSnapshotDetails
```

Id	CarParkName	CarParkLocation	CarParkCapacity	Lat	Lon
16	ABBEY	Northeast	359	53.349785	-6.254176

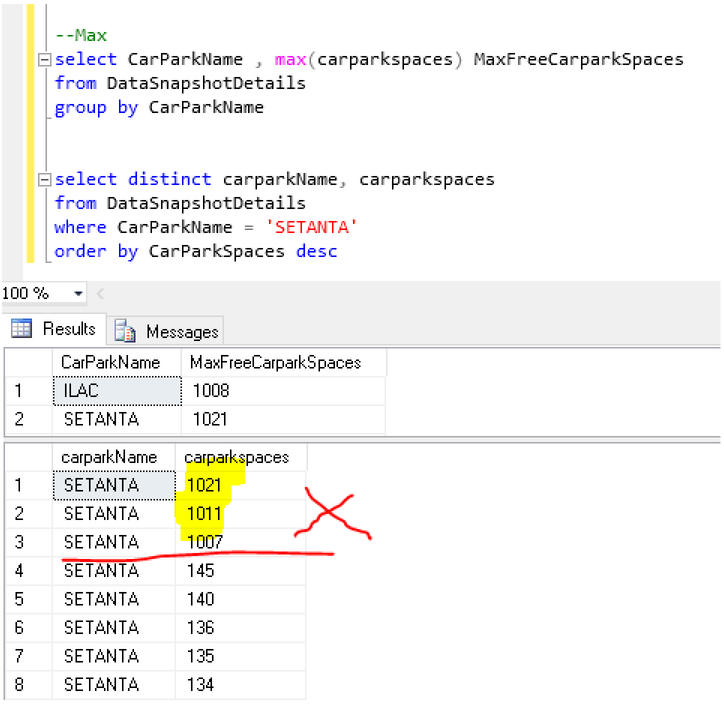
Id	SnapshotInternalReference	CarParkLocation	CarParkName	CarParkSpaces
17283	81DA9A0B-31BA-455F-8131-7F9C7E958136	Southeast	SETANTA	51

CarparkCapacity

I noticed this dataset does not contain the total capacity of the parkings, and I considered this was a mandatory value if I wanted to give any value to my visualizations. Manual work was done here to get this value by getting the MAX free spaces by parking and setting this value to the total capacity on a new CarParks table. This specific approach highlights a problem which will be tacked in the last section.

Removing records surpassing max capacity

While calculating this "max capacity" for each parking, I checked one by one to ensure the Max("carparkspaces") was not a wrong value, e.g. 999, and indeed I found some of this values appeared to be wrong in "SETANTA" parking:

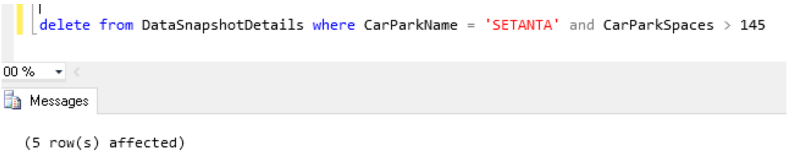


```sql
--Max
select CarParkName , max(carparkspaces) MaxFreeCarparkSpaces
from DataSnapshotDetails
group by CarParkName


select distinct carparkName, carparkspaces
from DataSnapshotDetails
where CarParkName = 'SETANTA'
order by CarParkSpaces desc
```

	CarParkName	MaxFreeCarparkSpaces
1	ILAC	1008
2	SETANTA	1021

	carparkName	carparkspaces
1	SETANTA	1021
2	SETANTA	1011
3	SETANTA	1007
4	SETANTA	145
5	SETANTA	140
6	SETANTA	136
7	SETANTA	135
8	SETANTA	134

I assumed these to be wrong and I removed them.



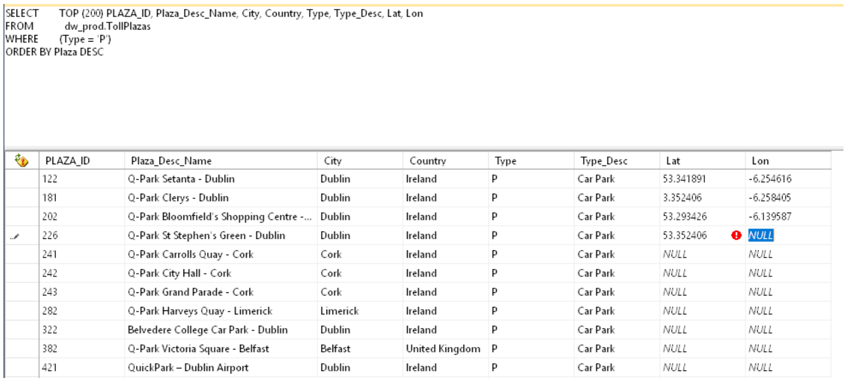
```
delete from DataSnapshotDetails where CarParkName = 'SETANTA' and CarParkSpaces > 145
```

```
(5 row(s) affected)
```

I repeated the same process for the rest of the parkings but the “MaxFreeCarparkSpaces” value seemed to be consistent across all the “*CarParkSpaces*” for that parking.

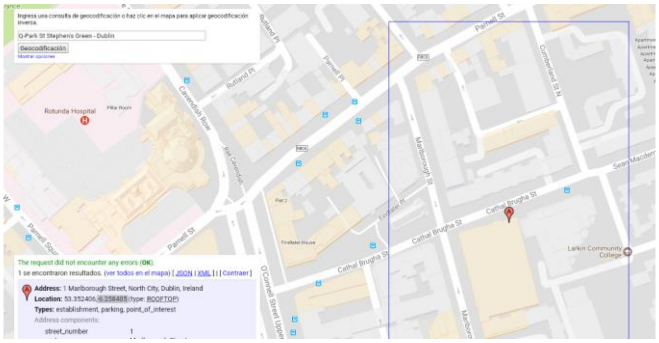
GeoLocating Datapoints

I also extracted geo locations for the Parkings through google geocoder tool. The screenshot below shows the work done for the 50 records of the discarded dataset in the first section above, but this task was also done for the 14 records of the DublinParking dataset.



```
SELECT      TOP (200) PLAZA_ID, Plaza_Desc_Name, City, Country, Type, Type_Desc, Lat, Lon
FROM          dw_prod.TollPlazas
WHERE       (Type = 'P')
ORDER BY Plaza DESC
```

PLAZA_ID	Plaza_Desc_Name	City	Country	Type	Type_Desc	Lat	Lon
122	Q-Park Setanta - Dublin	Dublin	Ireland	P	Car Park	53.341891	-6.254616
181	Q-Park Clerys - Dublin	Dublin	Ireland	P	Car Park	3.352406	-6.258405
202	Q-Park Bloomfield's Shopping Centre -...	Dublin	Ireland	P	Car Park	53.293426	-6.139587
226	Q-Park St Stephen's Green - Dublin	Dublin	Ireland	P	Car Park	53.352406	NULL
241	Q-Park Carrolls Quay - Cork	Cork	Ireland	P	Car Park	NULL	NULL
242	Q-Park City Hall - Cork	Cork	Ireland	P	Car Park	NULL	NULL
243	Q-Park Grand Parade - Cork	Cork	Ireland	P	Car Park	NULL	NULL
282	Q-Park Harveys Quay - Limerick	Limerick	Ireland	P	Car Park	NULL	NULL
322	Belvedere College Car Park - Dublin	Dublin	Ireland	P	Car Park	NULL	NULL
382	Q-Park Victoria Square - Belfast	Belfast	United Kingdom	P	Car Park	NULL	NULL
421	QuickPark – Dublin Airport	Dublin	Ireland	P	Car Park	NULL	NULL



Ingresa una consulta de geocodificación o haz clic en el mapa para aplicar geocodificación inversa

Q-Park St Stephen's Green - Dublin

The request did not encounter any errors (OK).

1 se encontraron resultados. (ver todos en el mapa) [JSON | XML] | [Contraer]

Address: 1 Marlborough Street, North City, Dublin, Ireland
Location: 53.352406,-6.258405 (type: ROOFTOP)
Types: establishment, parking, point_of_interest
Address components:
street_number 1

4.3. Filtering Stage

As seen above, rows have been removed as a part of a cleaning task, but for this specific phase no extra information was deleted.

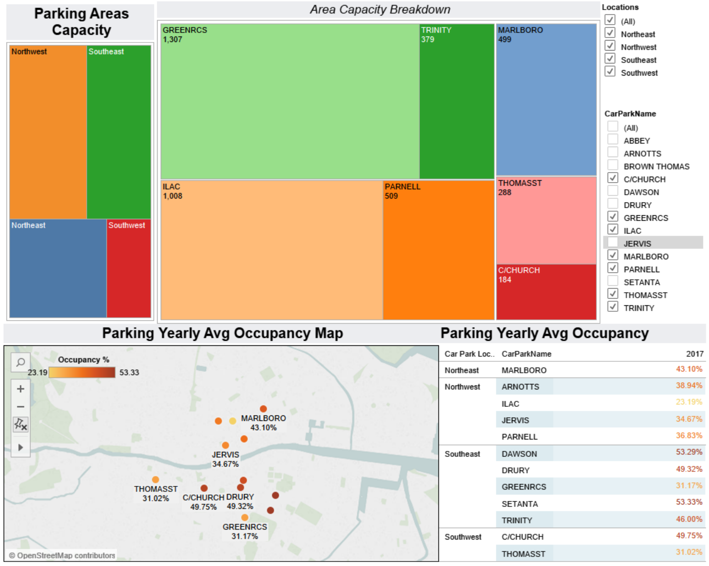
4.4. Mining Stage – Removing Extra rows

Same as in the previous point, besides the data that has been cleaned and removed, all the information contained in the dataset has been used.

4.5. Represent Stage

The solution contains several visualizations that provide an overall perspective of the parking sites, and offer the possibility of an interesting granularity which displays useful and meaningful data.

Main Dashboard



Locations
☑ (All)
☑ Northeast
☑ Northwest
☑ Southeast
☑ Southwest

CarParkName
☐ (All)
☐ ABBEY
☐ ARNOTTS
☐ BROWN THOMAS
☑ C/CHURCH
☐ DAWSON
☐ DRURY
☑ GREENRCS
☑ ILAC
☐ JERVIS
☑ MARLBORO
☑ PARNELL
☐ SETANTA
☑ THOMASST
☑ TRINITY

Parking Yearly Avg Occupancy

Car Park Loc..	CarParkName	2017
Northeast	MARLBORO	43.10%
Northwest	ARNOTTS	38.94%
	ILAC	23.19%
	JERVIS	34.67%
	PARNELL	36.83%
Southeast	DAWSON	53.29%
	DRURY	49.32%
	GREENRCS	31.17%
	SETANTA	53.33%
	TRINITY	46.00%
Southwest	C/CHURCH	49.75%
	THOMASST	31.02%

This dashboard contains 4 visualizations:

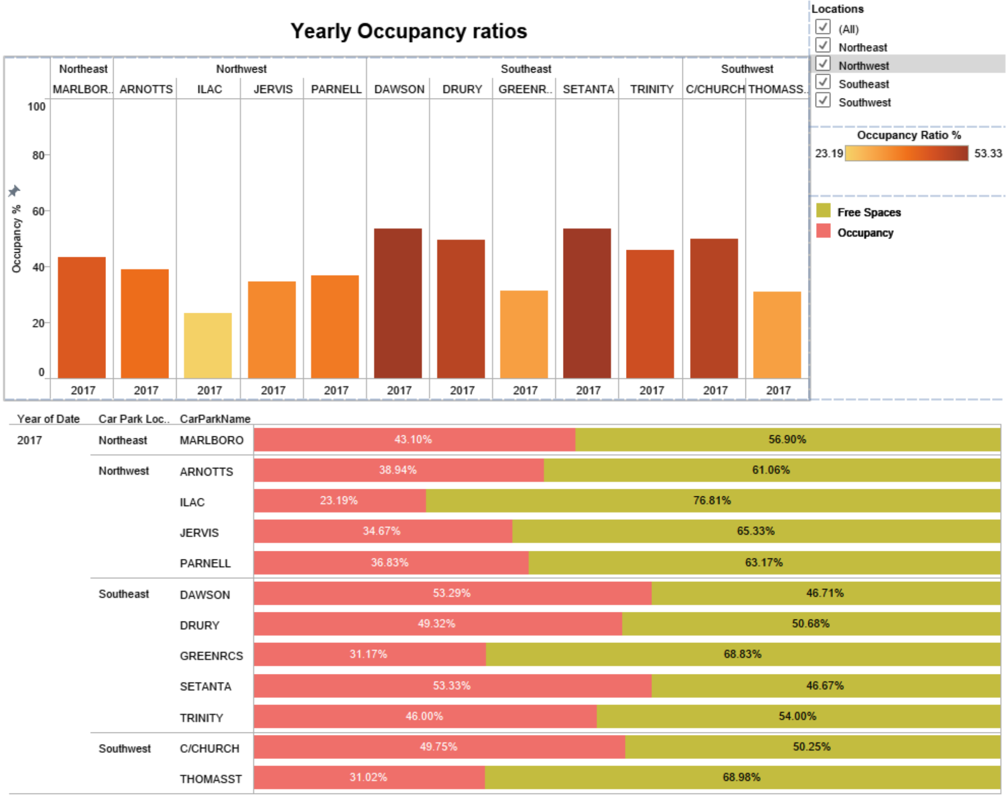
1. Parking Areas
   - ○ Showing sizes of the Areas.
2. Parking Capacity by Areas
   - ○ Displaying the capacity and other attributes, making easy to compare sizes between areas and parkings.
3. A map, displaying occupancy rates.

4. A table displaying similar information to the map but on a more readable way.

Parkings and areas above act as a filter for the rest of the elements.

Clicking on one of the Parkings of the map will bring us to the next visualization filtered by Area.

Yearly Occupancy Ratios

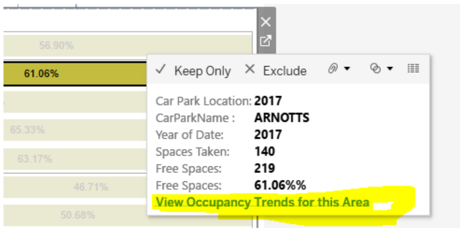


Year of Date	Car Park Loc..	CarParkName	Occupancy	Free Spaces
2017	Northeast	MARLBORO	43.10%	56.90%
	Northwest	ARNOTTS	38.94%	61.06%
		ILAC	23.19%	76.81%
		JERVIS	34.67%	65.33%
		PARNELL	36.83%	63.17%
	Southeast	DAWSON	53.29%	46.71%
		DRURY	49.32%	50.68%
		GREENRCS	31.17%	68.83%
		SETANTA	53.33%	46.67%
		TRINITY	46.00%	54.00%
	Southwest	C/CHURCH	49.75%	50.25%
		THOMASST	31.02%	68.98%

This is a second dashboard offering more detail about occupancy and free spaces. It makes easy to compare areas and parkings against each other.

The hover function highlights the selected information.

It also allows the option to keep on drilling down the info, this time to the Trends section:



Car Park Location: 2017
CarParkName : ARNOTTS
Year of Date: 2017
Spaces Taken: 140
Free Spaces: 219
Free Spaces: 61.06%%
View Occupancy Trends for this Area

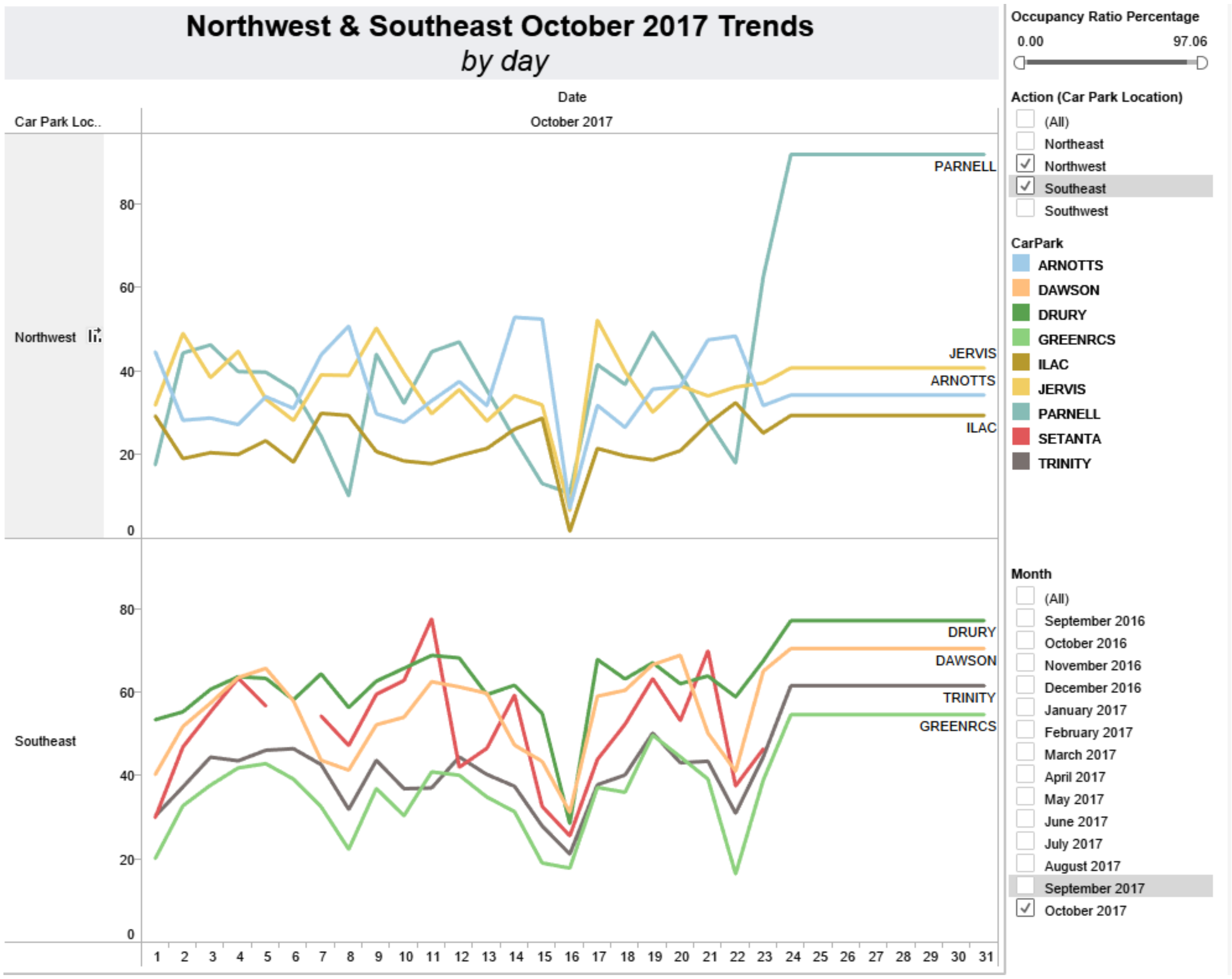
Occupancy Trends Dashboard

This third Dashboard has some useful information.

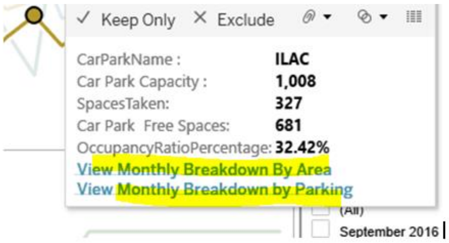
We can see here different trends by Area, on a monthly basis.

Several filters are available on the right side, including Occupancy Percentage, in case we want to focus in either peak, or off peak hours.

On the example below we can compare Northwest with Southeast for October, and see straight away which parkings were at highest capacity by hours.



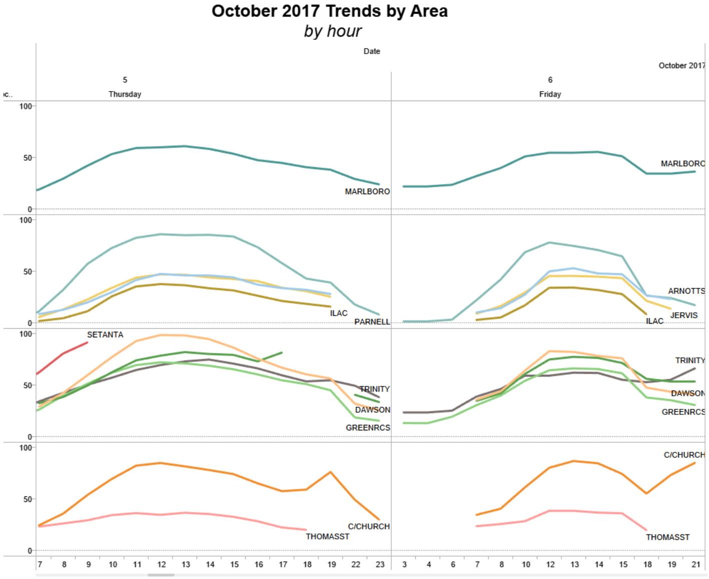
But, we want to see more. The menu gives us two more options that will be filtered by the selected area:



CarParkName : ILAC
Car Park Capacity : 1,008
SpacesTaken: 327
Car Park Free Spaces: 681
OccupancyRatioPercentage: 32.42%
View Monthly Breakdown By Area
View Monthly Breakdown by Parking

Monthly Breakdown By area

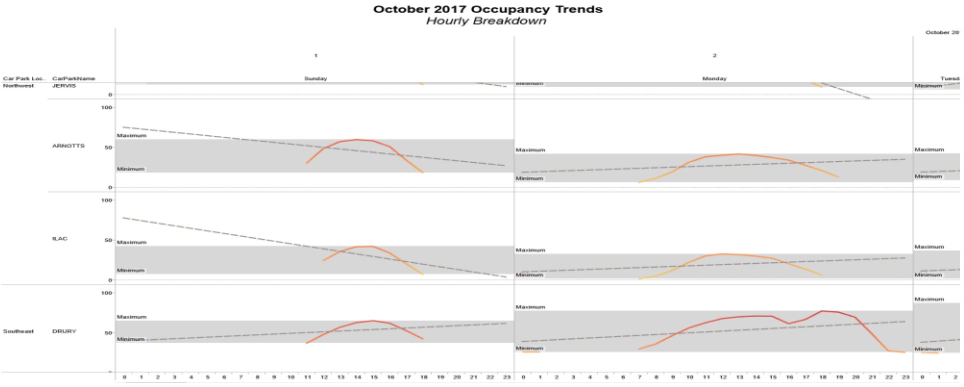
We can see the information in here grouped by areas hourly for each day of the month.



We can see how many parkings tend to have trends that we can compare against each other.

Monthly Breakdown by Parking

It may be the case where we want to have a clear picture of one of the Parkings and compare the data against itself, so we can go straight to it by selecting this visualization showing a clearer parking view, max and minimum values for each day along with trending lines.

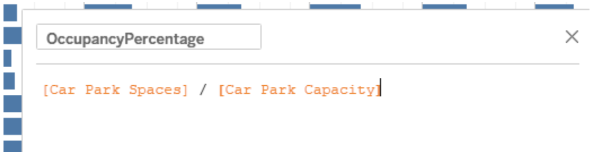


Most of these visualizations allow a filter by “*Occupancy*” so we can put our attention on peak or off peak hours.

4.6. Refining Stage

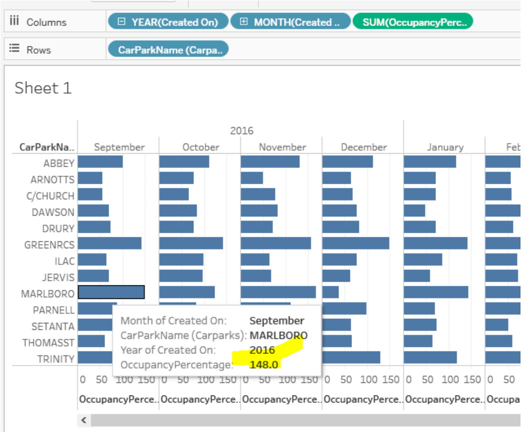
AVG occupancies over 100%

In this case I wanted to display percentages of occupancy. So a calculated field was created to represent this value (OccupancyPercentage), which is calculated based on the “*CarparkCapacity*” that had previously been manually added for each Parking.



```
[Car Park Spaces] / [Car Park Capacity]
```

But, ops! Something is wrong here. Some parkings show over 100% occupancy:



Month of Created On: September
CarParkName (Carparks): MARLBORO
Year of Created On: 2016
OccupancyPercentage: 148.0

Ran some queries to ensure my data was right, and found out the value displayed shouldn´t be a SUM, but an AVG.

Extra tooltips

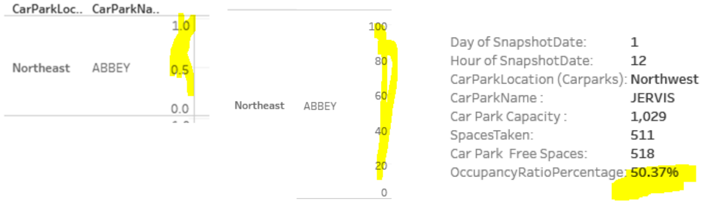
Longitudes and Latitudes have been removed from tooltips to minimize noise when reading stats.

Occupancy Ratio

“Car spaces” do not make any sense if not computed against other parkings. To do that, it has to be represented as a percentage within the parking total spaces. The label is formatted and shown as a percentage in all visualizations.

Ratio as Percentage

Eventually I noticed if I have a ratio not “percentaged” any axis based on this field will take values from 0.0 to 1, instead of 0 to 100. It does not make really much difference, but people wants to see numbers based on a 0-100 scale. The solution was to *percentage* this calculated field, replace the default “percentage formatting” and edit the labels to add a “%” symbol on them.



Day of SnapshotDate: 1
Hour of SnapshotDate: 12
CarParkLocation (Carparks): Northwest
CarParkName : JERVIS
Car Park Capacity : 1,029
SpacesTaken: 511
Car Park Free Spaces: 518
OccupancyRatioPercentage: 50.37%

More calculated fields

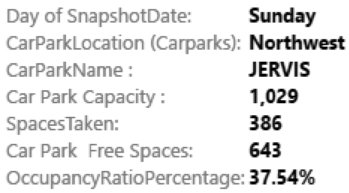
Several calculated fields added:

- FreeSpaces

- FreeRatioPercentage
- TakenSpaces
- OccucpancyRatioPercentage

Tooltips must show valuable information

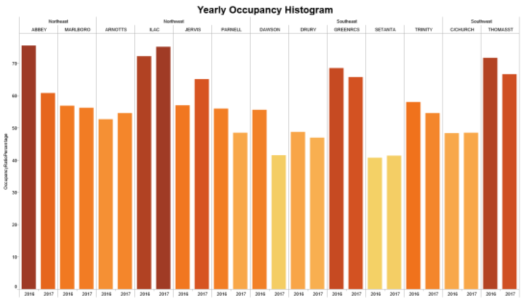
It is not just the visualization, but the information a specific point gives us:



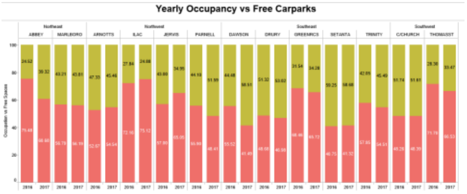
Day of SnapshotDate:	**Sunday**
CarParkLocation (Carparks):	**Northwest**
CarParkName :	**JERVIS**
Car Park Capacity :	**1,029**
SpacesTaken:	**386**
Car Park Free Spaces:	**643**
OccupancyRatioPercentage:	**37.54%**

A graph missing something

In the graph below I wanted to show stats of occupancy. But although it gives a very good graphical representation of the data, there was something missing.

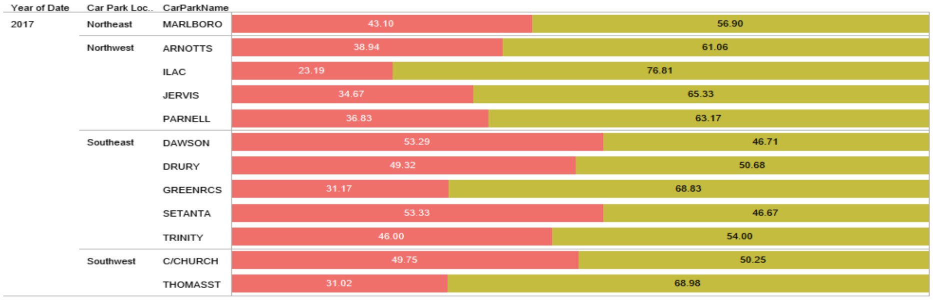


In this case I opted for a combined bar that would show free ratio vs occupancy ratio percentages as shown below.

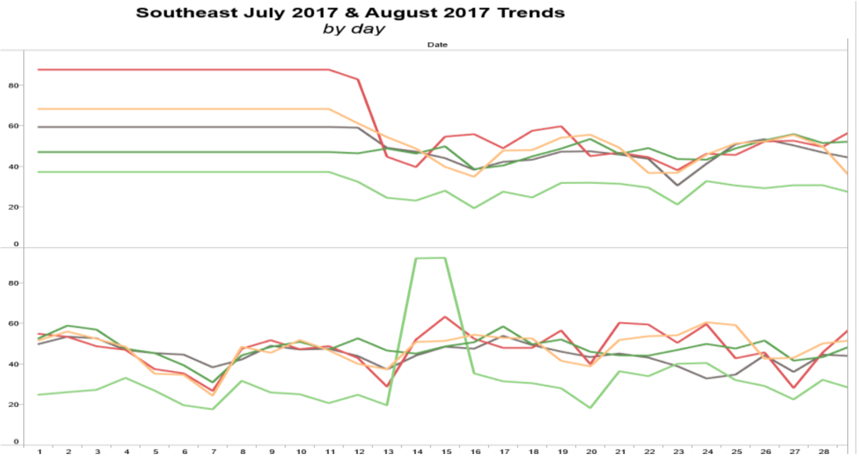


Swapping rows per columns

Obviously, I ran into a readability issue for the last image above. When I was refining my views, I flipped the bars (see OccupancyBarsDh dashboard):

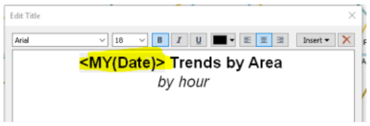


Another example for this scenario is the Occupancy Trends By Area. By moving the Month to the columns side, we have a clearer visualization, so data can be compared at first glance, rather than having to scroll to compare different months:



Accurate info on Titles

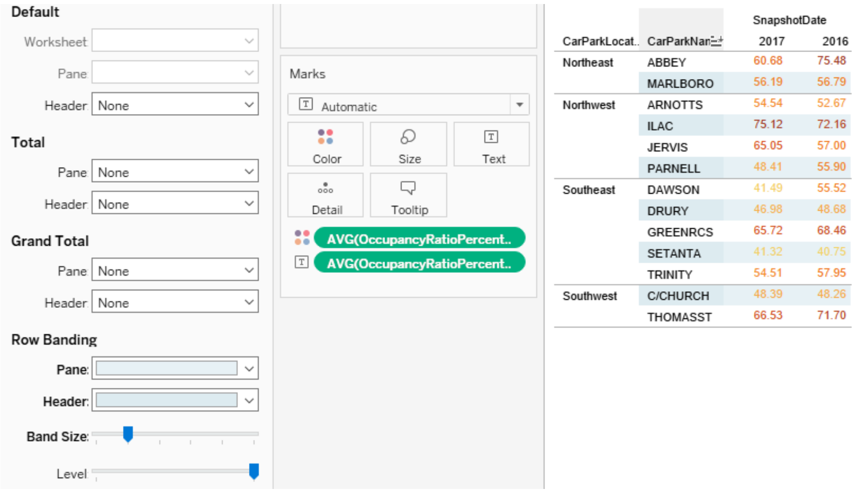
Special attention has been put on titles displaying filters.



<MY(Date)> Trends by Area
by hour

Contrast

When using “Gold” coloured palettes, the yellow values did not have enough contrast with a white background, so I had to modify the cell background colour so it would be readable.



Default

Worksheet

Pane

Header: None

Total

Pane: None

Header: None

Grand Total

Pane: None

Header: None

Row Banding

Pane:

Header:

Band Size:

Level

CarParkLocat..	CarParkNam	SnapshotDate 2017	 2016
Northeast	ABBEY	60.68	75.48
	MARLBORO	56.19	56.79
Northwest	ARNOTTS	54.54	52.67
	ILAC	75.12	72.16
	JERVIS	65.05	57.00
	PARNELL	48.41	55.90
Southeast	DAWSON	41.49	55.52
	DRURY	46.98	48.68
	GREENRCS	65.72	68.46
	SETANTA	41.32	40.75
	TRINITY	54.51	57.95
Southwest	C/CHURCH	48.39	48.26
	THOMASST	66.53	71.70

4.7. Interact Stage

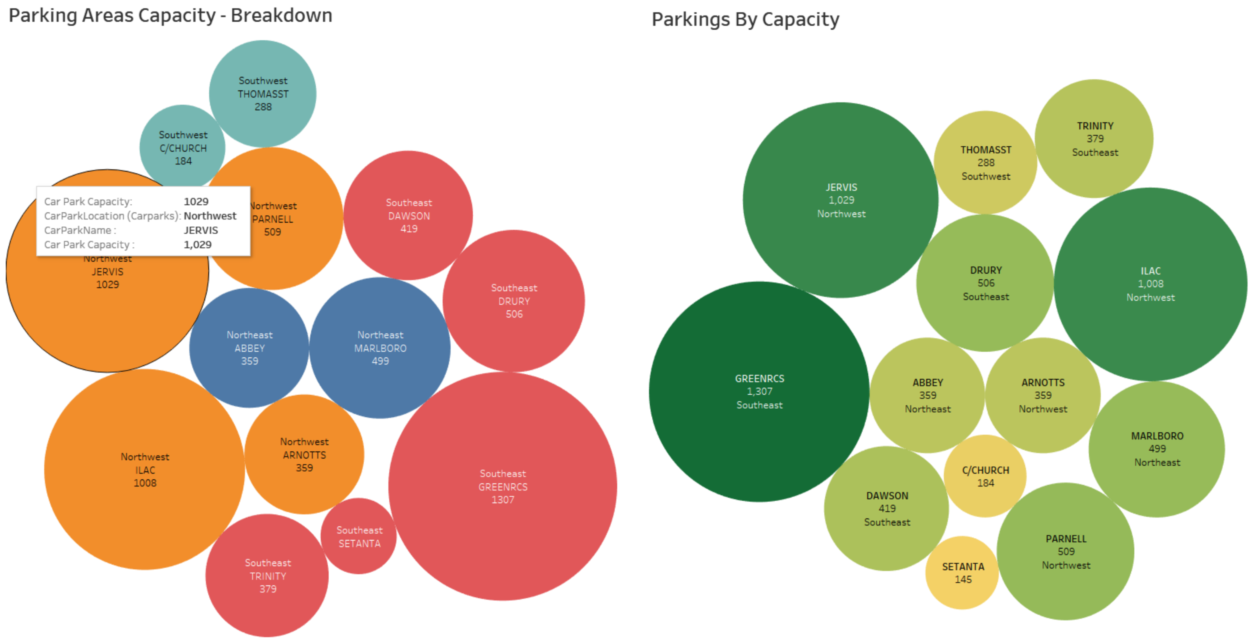
As detailed above, some interactions so the user can drill into the data have been provided.

- Several filters solving the peak / off peak hours
- Actions filtering current and targeted visualizations.
- Contextual menus allowing navigating between visualizations within selected parking areas or dates.
- Highlight actions for better readability.

5. Problems and Solutions

Different bubble sorts for same data

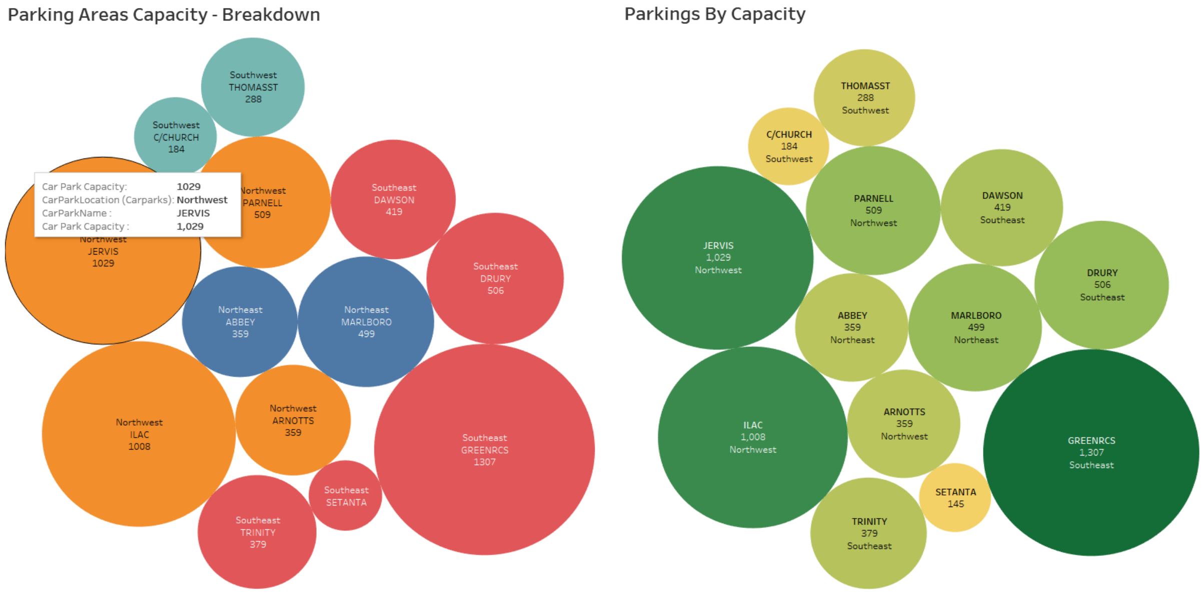
I wanted to give a visual comparison of capacity between parkings and somehow the first way wasn´t really accomplishing what I wanted to show. So I decided to create the same visualization with different colours, which would attract the attention where I wanted to be. I found out then the “Packed Bubbles” visualization was displaying data in different positions. And this was altering the line of the story, as it looks at first glance (by colour) that we are looking to different data.


Parking Areas Capacity - Breakdown

Parkings By Capacity

Image shows same data displayed in different positions

This was a problem for what I was trying to achieve. I investigated online, and came out with different tricks, such as here. After doing some tests, I came up with the sorting that was altering the positions of the bubbles. Although the 2 images are trying to show same information from different perspectives, having the bubbles in same positions makes it easier to come up with a relationship between two pictures.


Parking Areas Capacity - Breakdown

Parkings By Capacity

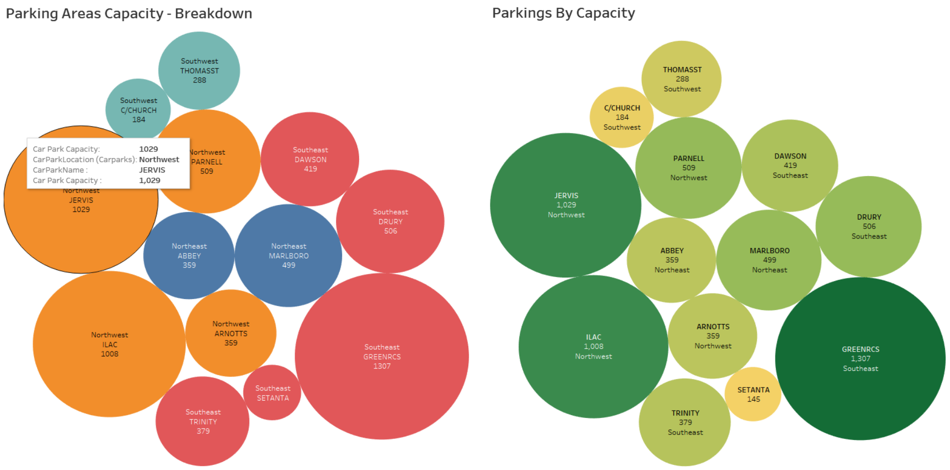
Image shows same same representation of the data, coloured by different factors. Left is coloured by Areas, and right is coloured by Capacity.

It is easy to understand these two pictures represent same data with different colour schemes drawing the attention to different aspects of the visualization.

Bubbles vs Treemaps

After working some more time, I came to the point where I realized I was trying to show something in two similar visualizations, and this was an indication something was wrong.

I started digging a bit and came to the conclusion that I was using the wrong chart. These two visualizations below were replaced by the treemap below.


Parking Areas Capacity - Breakdown

Parkings By Capacity

Bubbles try to give a visual size representation of parkings , first within a parking area, and secondly relative to other parkings.

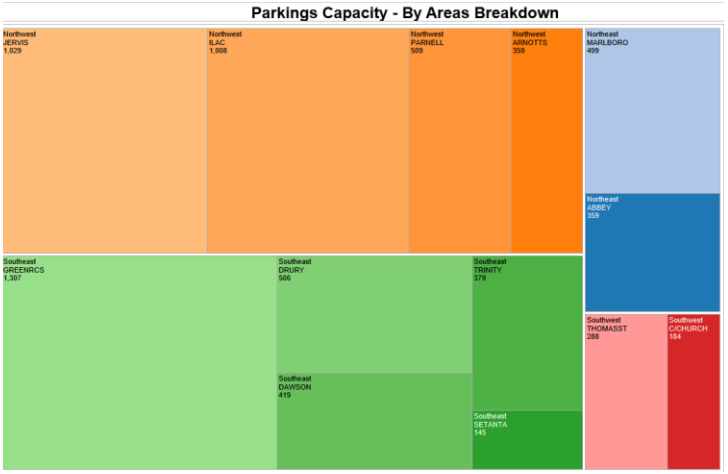
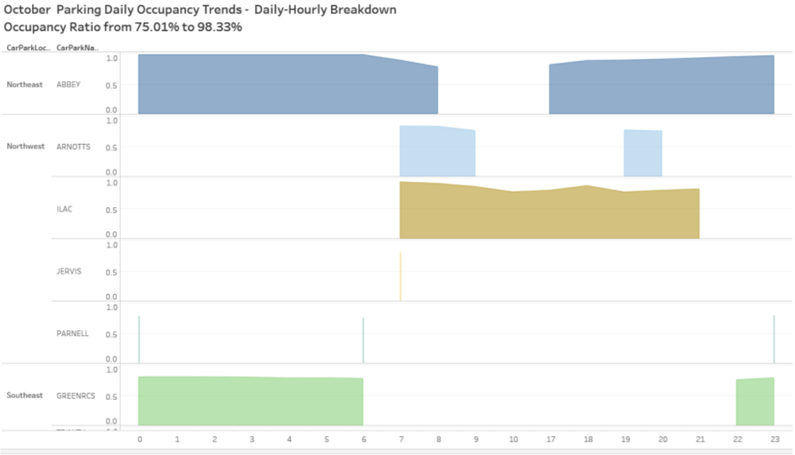


Image above solves the problem with an easy graphical representation.

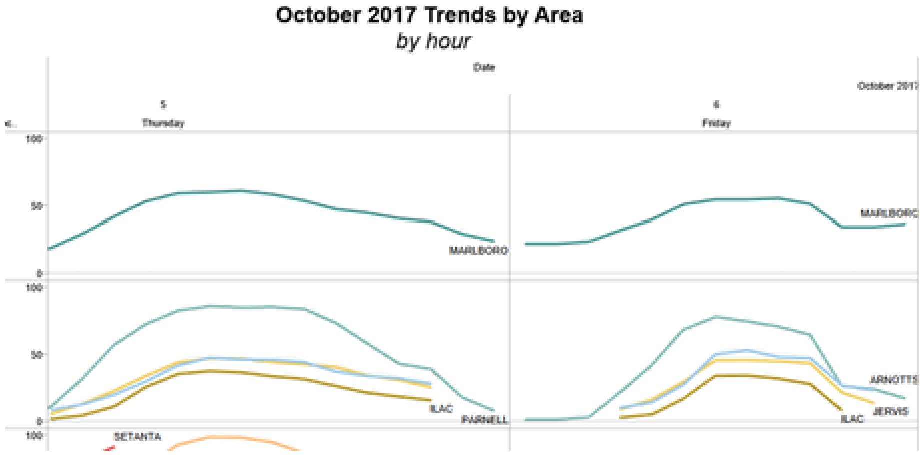
Comparing data together

I was trying to represent meaningful and valuable range hours so we could tackle companies or individuals, either from a peak time perspectives, parking offers, etc. or on low times. E.g. hiring empty spaces and promoting them and lower prices.



October Parking Daily Occupancy Trends - Daily-Hourly Breakdown
Occupancy Ratio from 75.01% to 98.33%

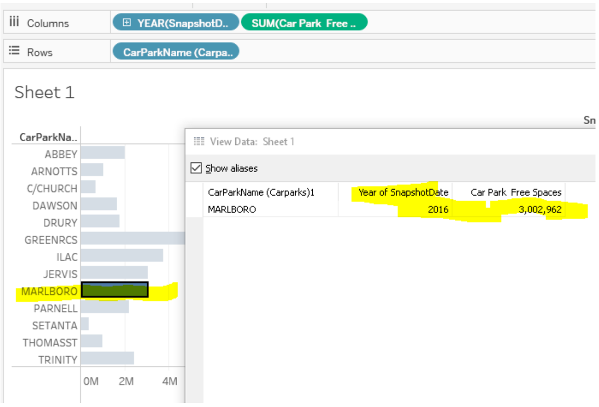
This visualization above turned into the one below. The information is much clearer, and the “OccupancyRatio” filter solves the problem of the data we want to see:



October 2017 Trends by Area
by hour

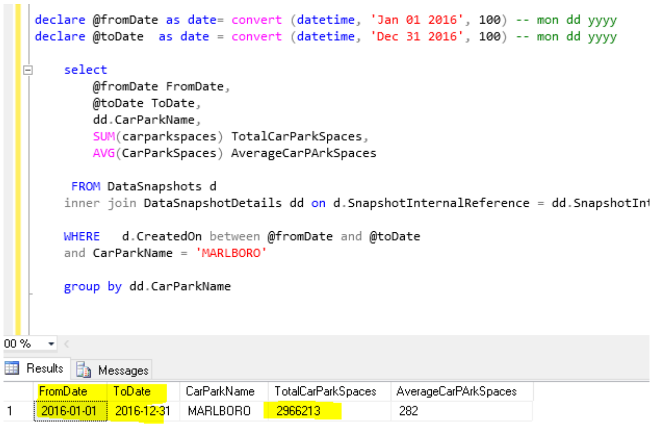
Data calculations not matching

I ran into an issue that took me few hours to resolve. I simplified the visualization so I could get to the bottom of the problem. For some reasons, total calculations were not working when segmenting the data by date. In the screenshot below the mismatch between Tableau and SQL db can be appreciated:



☑ Show aliases

CarParkName (Carparks)1	Year of SnapshotDate	Car Park Free Spaces
MARLBORO	2016	3,002,962



```
declare @fromDate as date= convert (datetime, 'Jan 01 2016', 100) -- mon dd yyyy
declare @toDate  as date = convert (datetime, 'Dec 31 2016', 100) -- mon dd yyyy

    select
        @fromDate FromDate,
        @toDate ToDate,
        dd.CarParkName,
        SUM(carparkspaces) TotalCarParkSpaces,
        AVG(CarParkSpaces) AverageCarPArkSpaces

     FROM DataSnapshots d
    inner join DataSnapshotDetails dd on d.SnapshotInternalReference = dd.SnapshotInt

    WHERE   d.CreatedOn between @fromDate and @toDate
    and CarParkName = 'MARLBORO'

    group by dd.CarParkName
```

	FromDate	ToDate	CarParkName	TotalCarParkSpaces	AverageCarPArkSpaces
1	2016-01-01	2016-12-31	MARLBORO	2966213	282

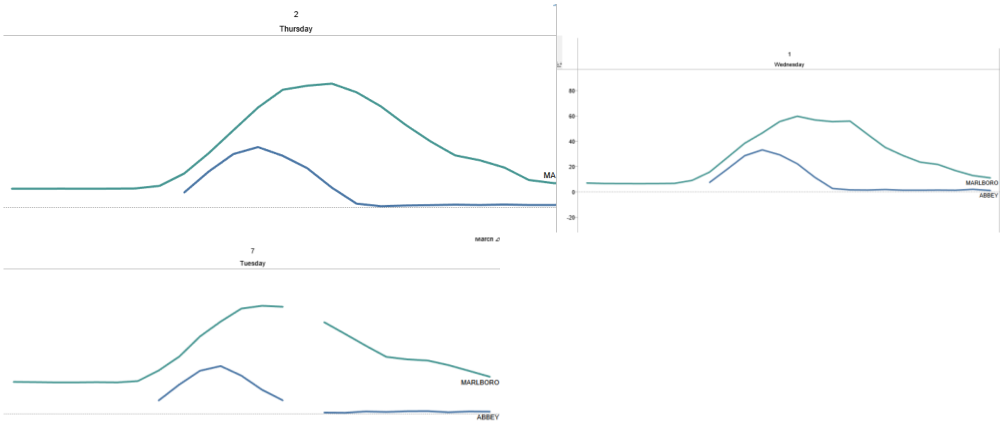
2017 calculations were matching perfectly, so I assumed Tableau was doing some internal computations. This would require extensive investigation.

7. Conclusion

One of the most interesting things we find is the existence of trends. From a business perspective we would be interested in both, peak and off peak hours.

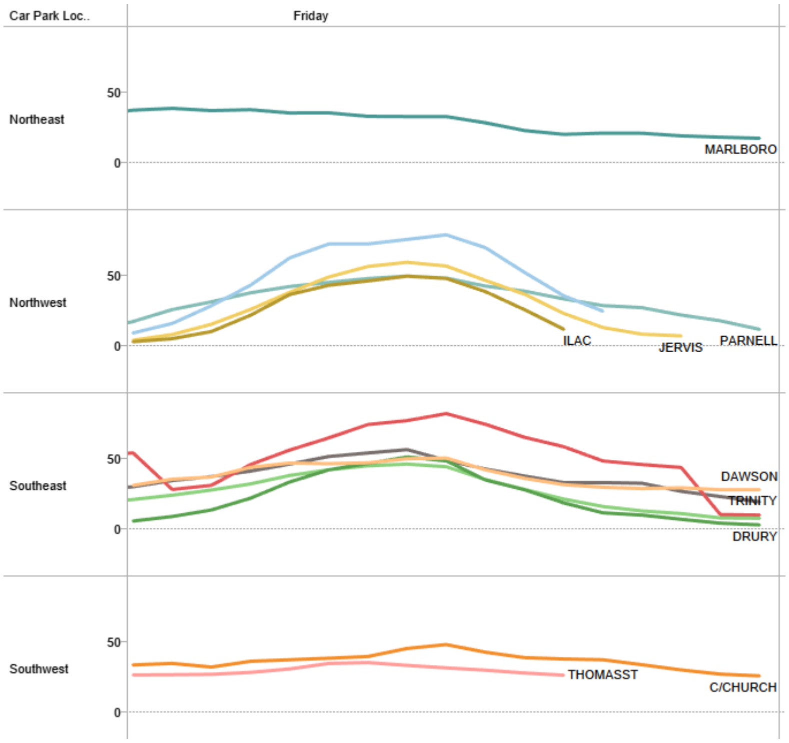
The presence of trends brings up valuable information that indicates possibility of predictions. Peak and off peak times are clearly identified within a matter of seconds.

For example, the three screenshots below are showing occupancy trends for 2 parkings on Northeast area.



We could be able of determining the number of free car spaces by hour. This could give us the ability to tackle parking companies and offer lower rates to customers, or special offers based on hour windows.

It is also interesting the fact that parkings from same areas would have similar curves, as shown below



This may be an indication of saturated areas meaning there is a lack of street space. As a consequence all parkings on the area suffer high peaks of demand at specific hours.

As a note, it is **very important** to remark that all these ratios have been calculated based on **fixed car park spaces.** This means, there is no evolution on this value along the history. Modifications on this value could affect percentages.

So as per conclusions we can say:

- There are specific defined trends across days of week for each parking. E.g. some days show remarkable peak hours while a Sunday shows a very low activity.

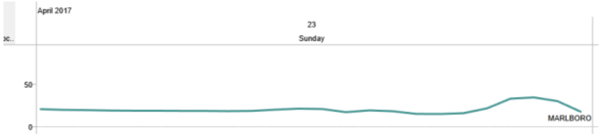


Image shows Sunday 23 Marlboro occupancy rate by hours. The hours scale is left to right 00 to 23h (not shown on picture)

- These trends may be an indication of predictors of available parking spaces based on day, and hour.
- Considerable visual resemblance within parkings on the same areas, which may be an indication of congestion in some locations.
- None of the parkings reaches 100% of occupancy.

Further Analysis areas recommendation:

- Better modelling of the tables recommended. CarParkSpaces has to be linked with dates as well, so there is a history of this value that can be linked with each record.
- Prediction of trends
- Crossing this data with hourly rates could easily come up with benefit vs costs visualizations.
- Similar applications to internal business model can easily be performed.

Business actions based on data:

- Approach to parkings recommended. Special offers based on times strategy.