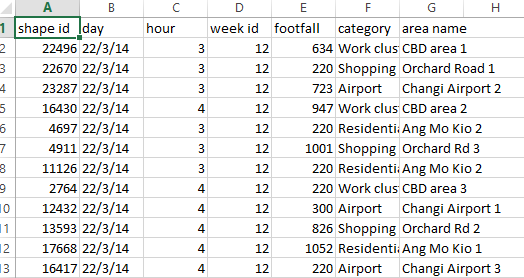
Here’s what the IDA Footfall Footfall count by hour by day for specific locations

Sample data-set look like:



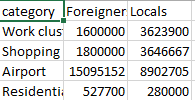
It has

|  |  |
| --- | --- |
| **Field** | **Value** |
| Shape id | Shape id |
| day | Date of the specific time |
| Hour | Hours taken for each footfall count |
| Week id | Week id |
| Footfall | the number of people entering a shop or shopping area in a given time. |
| Category | Includes:   1. Work cluster 2. Shopping cluster 3. Airport 4. Residential Areas |
| Area name | 1. Work Cluster: CBD Area 1, CBD Area 2, CBD Area 3 2. Shopping Cluster: Orchard Road 1, Orchard 2, Orchard 3 3. Airport: Changi Airport 1, Changi Airport 2, Changi Airport 3 4. Residential Area: Ang Mo Kio 1, Ang Mo Kio 2 |

So, if you can see that the area name are all around the same area, for example, work cluster are all at central areas(CBD area 1,2 etc) and Shopping cluster are all at Orchard Rd….

So, instead of finding a dataset that can link with the area name, I find another dataset based on **category** instead.

I have found another dataset on overall estimation footfall count of foreigners vs locals within **2014** in Singapore based on the **category** from IDA dataset. Here’s what my dataset looks like:

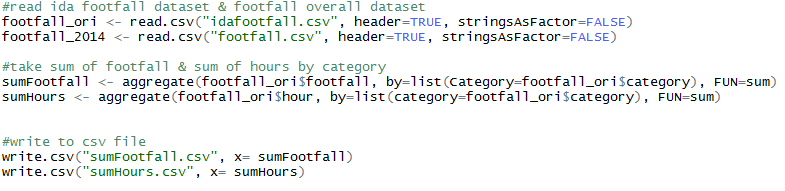


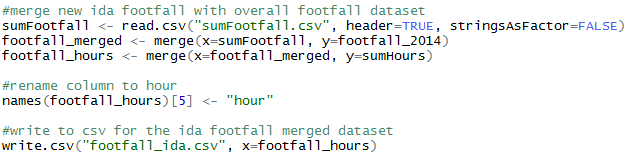
|  |  |
| --- | --- |
| **Field** | **Value** |
| category | Includes:   1. Work cluster 2. Shopping cluster 3. Airport 4. Residential Areas |
| Foreigners | 1. Work cluster: Footfall count of foreigners in central areas(CBD area) 2. Shopping cluster: Footfall count of foreigners in Orchard Rd 3. Airport: Footfall count of foreigners visiting Singapore 4. Residential areas: PRs(foreigners) living in Singapore |
| Locals(Mainly Permanent Residents(PRs) and Citizens) | 1. Work cluster: Footfall count of locals in central areas(CBD area) 2. Shopping cluster: Footfall count of locals in Orchard Rd 3. Airport: Footfall count of outbound Singapore 4. Residential areas: Residents living in Ang Mo Kio |

*Here’s where I found the datasets(if you want to know how):* [*https://www.dropbox.com/s/74iyrs2xk44ni28/FootfallDatasets.docx?dl=0*](https://www.dropbox.com/s/74iyrs2xk44ni28/FootfallDatasets.docx?dl=0)

After getting the datasets I wanted, I want to merge them based on the category but first, I have to take the sum of the **footfall count** and **hours** from the IDA dataset based on the **category.**

To do that, in R:





After merging, I’m ready to **further analyze,**

**Ideally,** my plan of analyzing are:

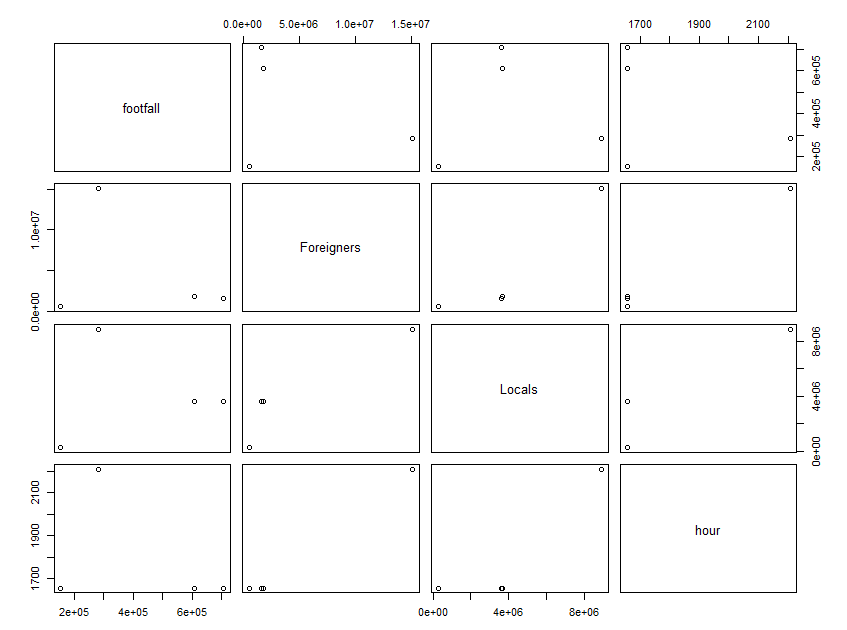
1. Get another dataset from the internet which is “the overall footfall estimation count in 2014” dataset
2. Merge them with the IDA dataset based on **category**
3. Plot matrix correlation of the merged datasets to see its linearity
4. Check the correlation of all pair-wise to see what other attributes I can also consider looking at
5. Run regression with having **footfall** being the independent variable and dependent variables are **foreigners** and **locals** to check if they are statically significant.

But the main problem is: both datasets are only within **2014** which is not surprising if the scatterplots are not linear and p-value not being statically significant.

Here’s what I tried:

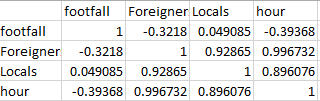


Output:



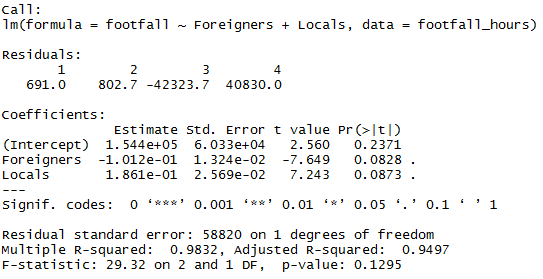


Output:





Output:



So, I was thinking **if** the full records dataset from IDA consists of a decade records, it wouldn’t be a problem to merge based on **year** and **category?**