NYC 311 Complaints data analysis

Big data analysis C.Sc. 84030 : Final project

VISHAL BHARTI

311 Service

- 311 is New York City's main source of government information and non-emergency services.
- NYC receives 311 calls for non-emergency services from it's residents, businesses and visitors.
- Response time for these calls is longer than those for emergency (911) calls.
- The data is open and updated frequently.
- Data used in the project 1/1/2010-12/7/2016 (~8.8 gb).
- Comprise of 14.2 million rows and 53 columns.

Hypothesis

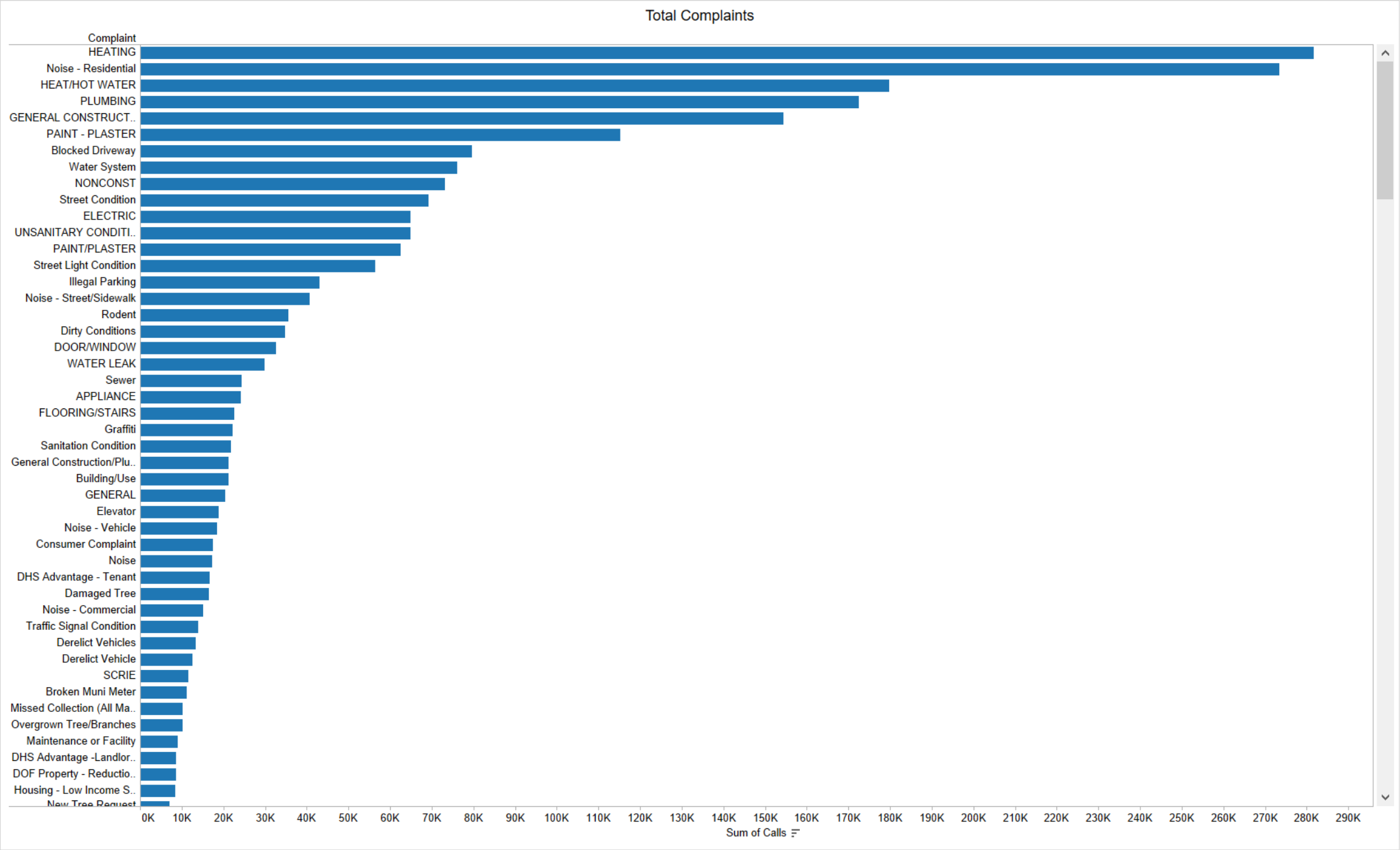
- Complaints will be vary for each location.
- Complaints will vary based on time of day, the day of week and time of the year.
- Noise complaints will be higher in residential areas.
- Parking violations complaints will increase during weekends.
- Street light conditions complaints will be higher during night time.
- Correlation between the number of complaints and response time.

Methodology

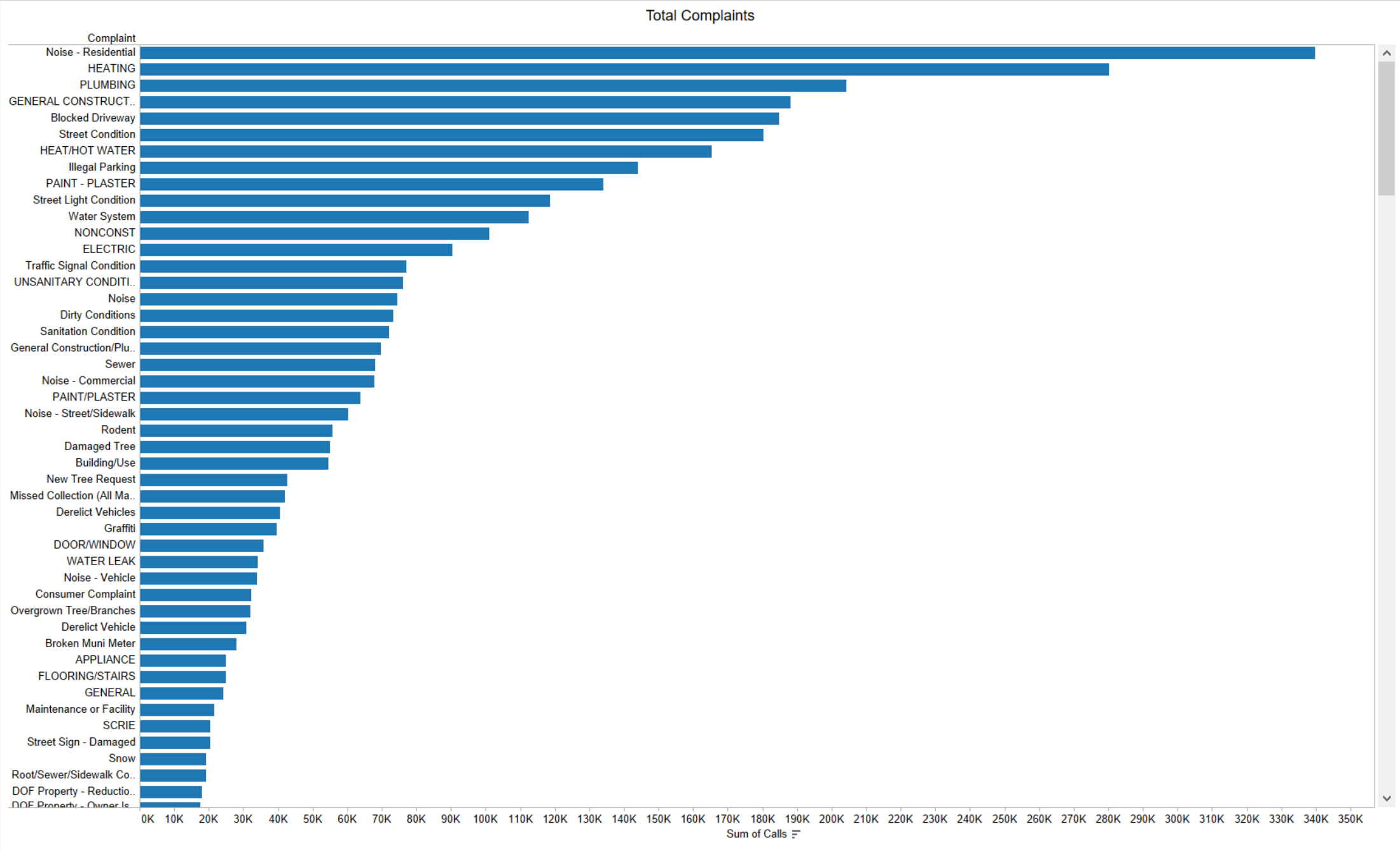
- Filtered dataset to remove noisy records. Some records had missing or incorrect closing date.
- Used Pyspark to filter the data based on keys.
- The data size was reduced from 8.8 gb to approx. 300mb.
- Visual analytics was used to draw information from the resulting dataset.
- Tableau was used for visual analytics.

Complaints by location

- For Bronx, ‘Noise’, ‘Heating’ and ‘Street Light Conditions’ were the most common complaints.
- For Brooklyn, ‘Noise’, ‘Heating’ and ‘Street Conditions’ were the most common complaints.
- Manhattan also had, ‘Noise’, ‘Heating’ and ‘Street Conditions’ as the most reported complaints.
- Queens had ‘Noise’, ‘Street Light Conditions’ and ‘Blocked Driveway’ as the most reported complaints.
- Staten Island had ‘Street Light Conditions’, ‘Street Conditions’ and ‘Water System’ as the most reported complaints.
- ‘Noise’, ‘Heating’ and ‘Street Light Conditions’ were the overall most reported complaints.



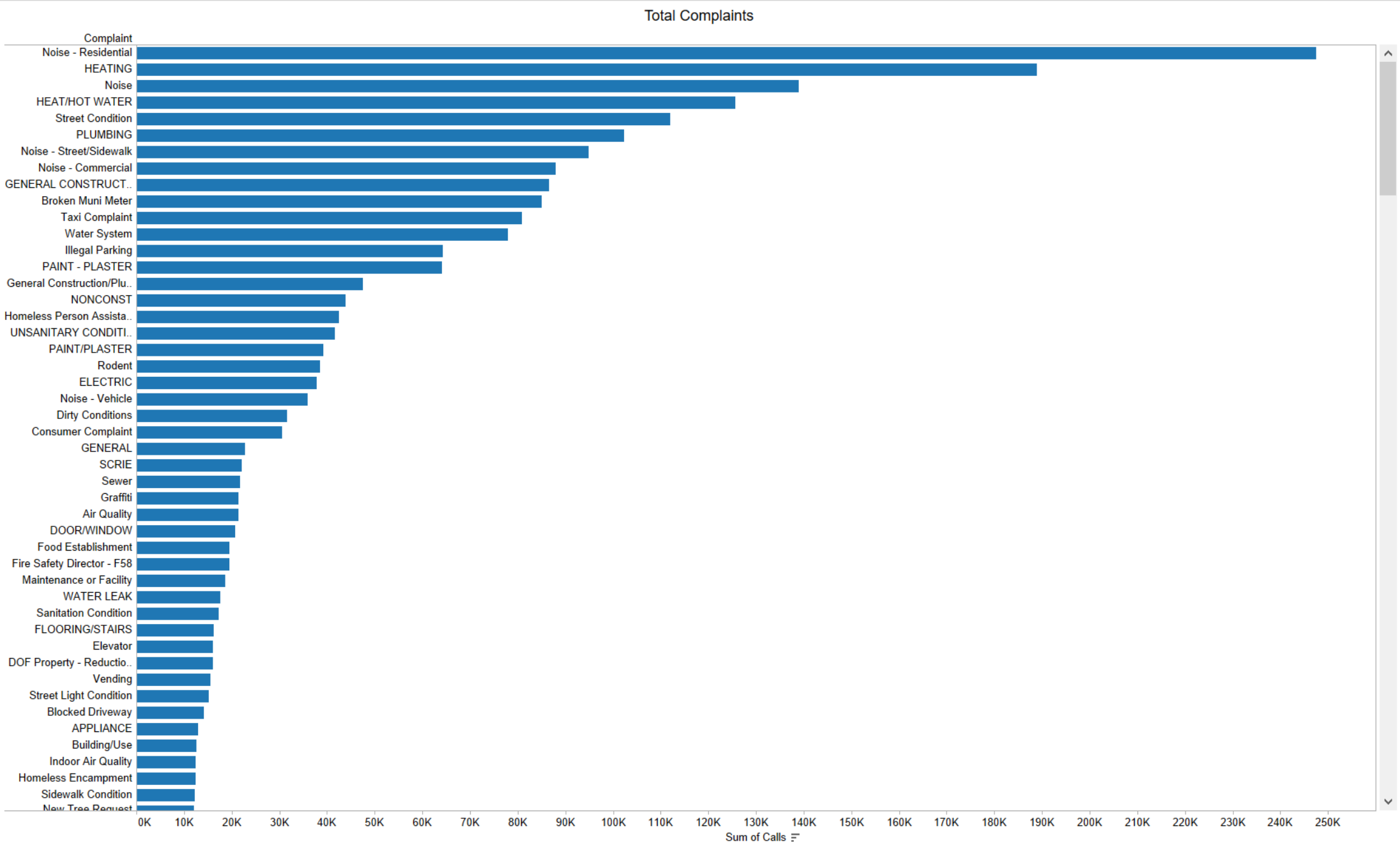
Map | TreeMap | Complaints TreeMap | Comp Dash | Complaint by hour | Average Response time | Complaints by month | **Total Complaints**



Borough

- (All)
- BRONX
- BROOKLYN
- MANHATTAN
- QUEENS
- STATEN ISLAND

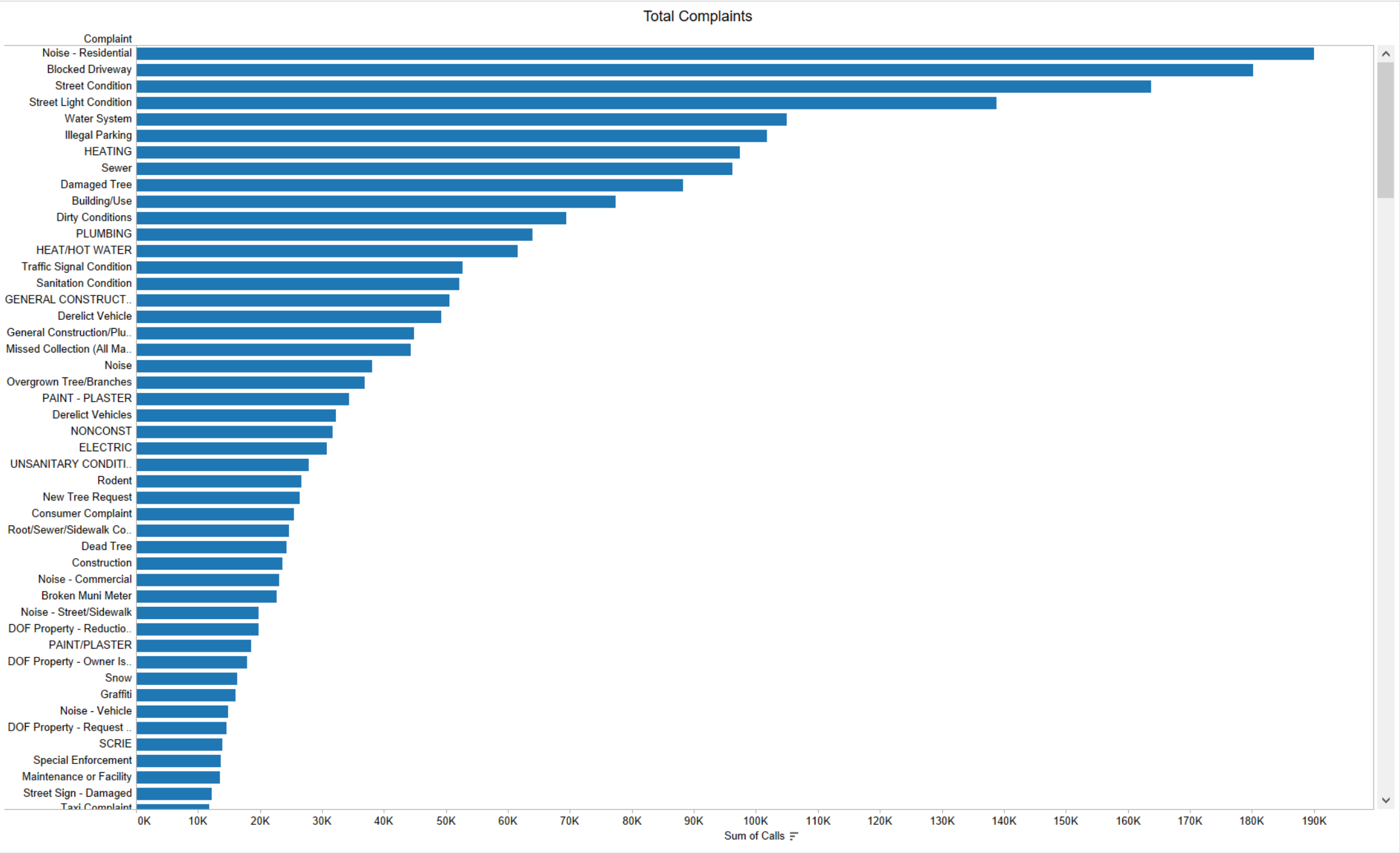
Map | TreeMap | Complaints TreeMap | Comp Dash | Complaint by hour | Average Response time | Complaints by month | **Total Complaints**



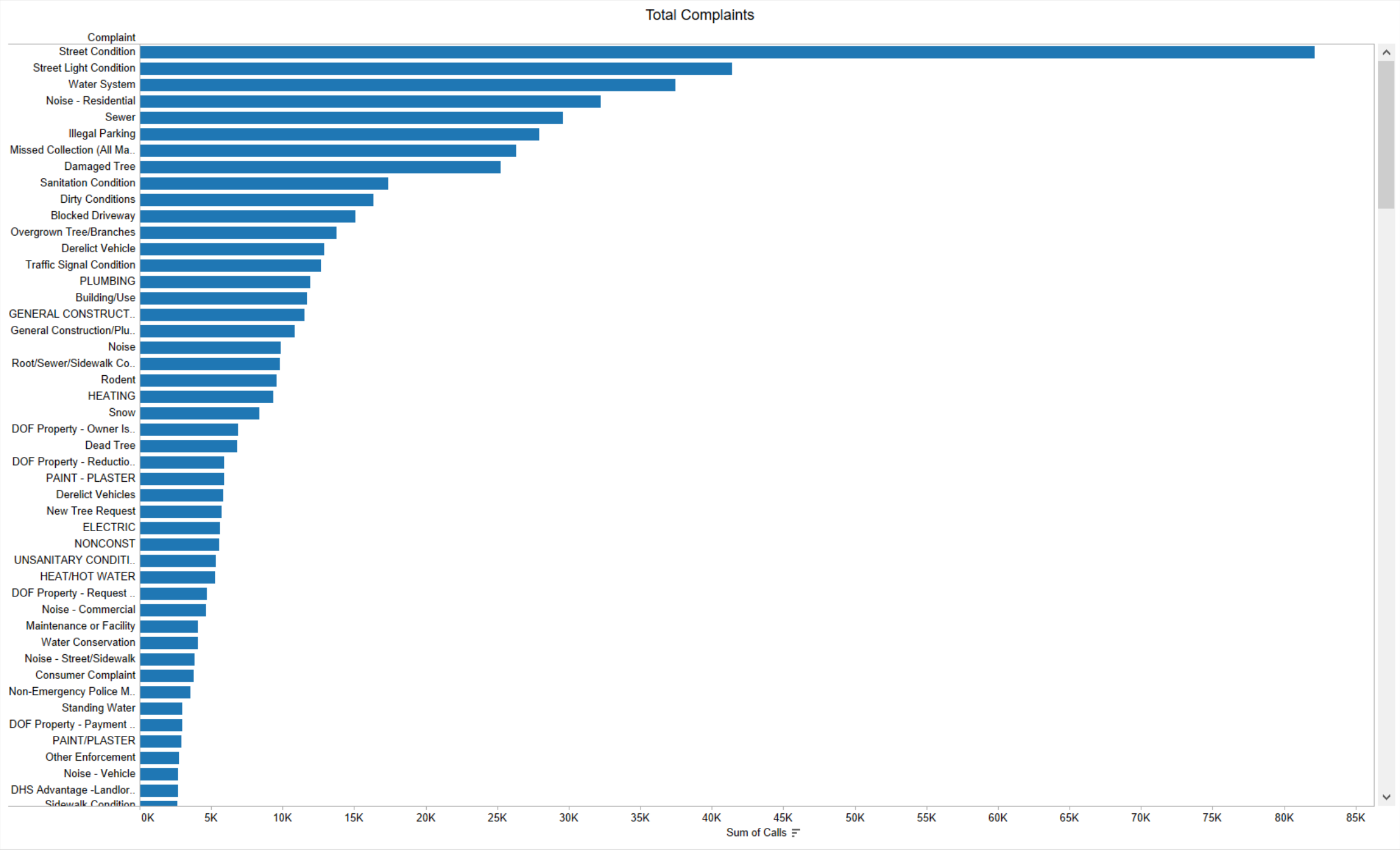
Borough

- (All)
- BRONX
- BROOKLYN
- MANHATTAN
- QUEENS
- STATEN ISLAND

Map | TreeMap | Complaints TreeMap | Comp Dash | Complaint by hour | Average Response time | Complaints by month | **Total Complaints**



Map | TreeMap | Complaints TreeMap | Comp Dash | Complaint by hour | Average Response time | Complaints by month | **Total Complaints**



Borough

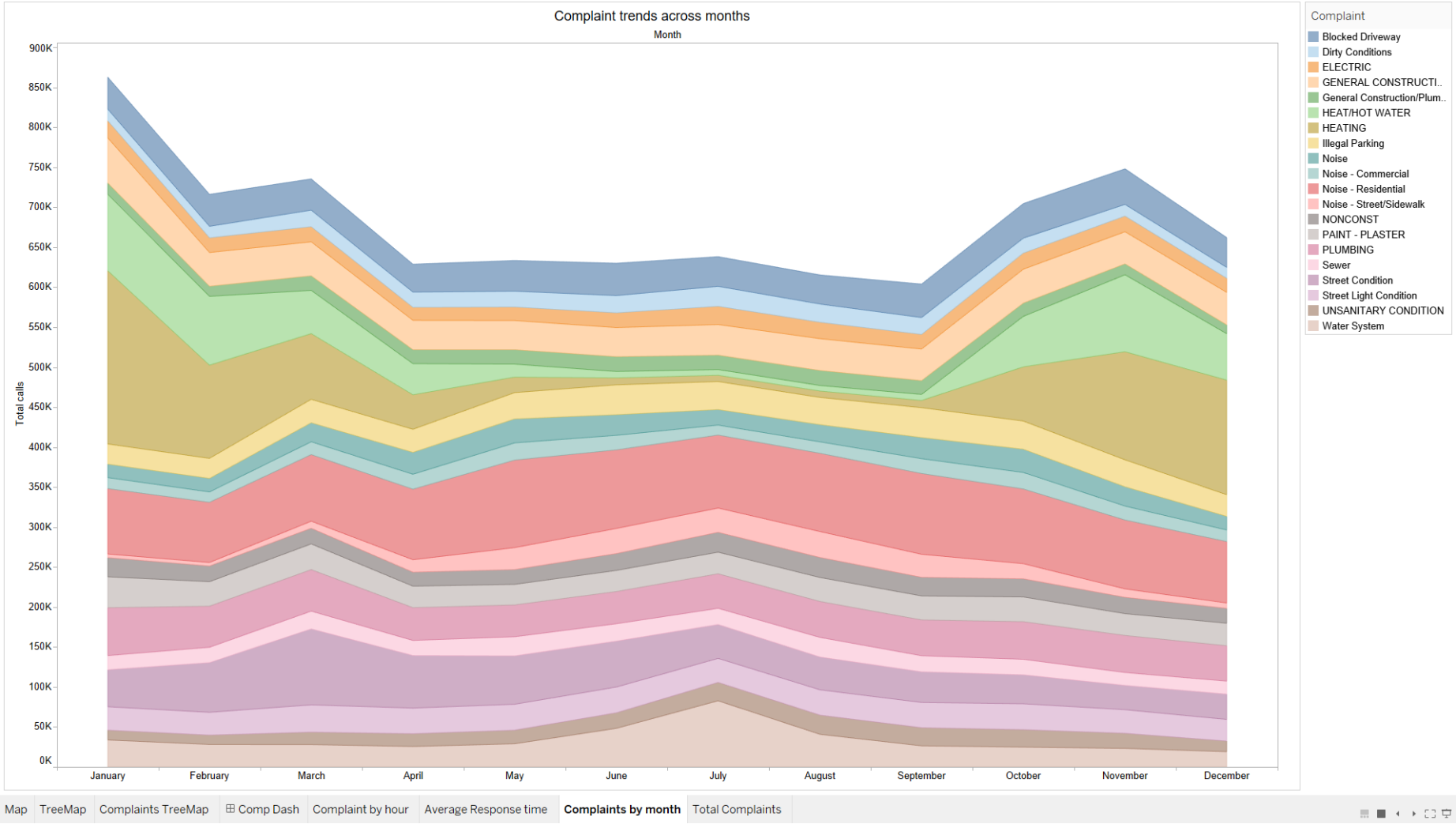
- (All)
- BRONX
- BROOKLYN
- MANHATTAN
- QUEENS
- STATEN ISLAND

Map | TreeMap | Complaints TreeMap | Comp Dash | Complaint by hour | Average Response time | Complaints by month | **Total Complaints**

Complaints by month

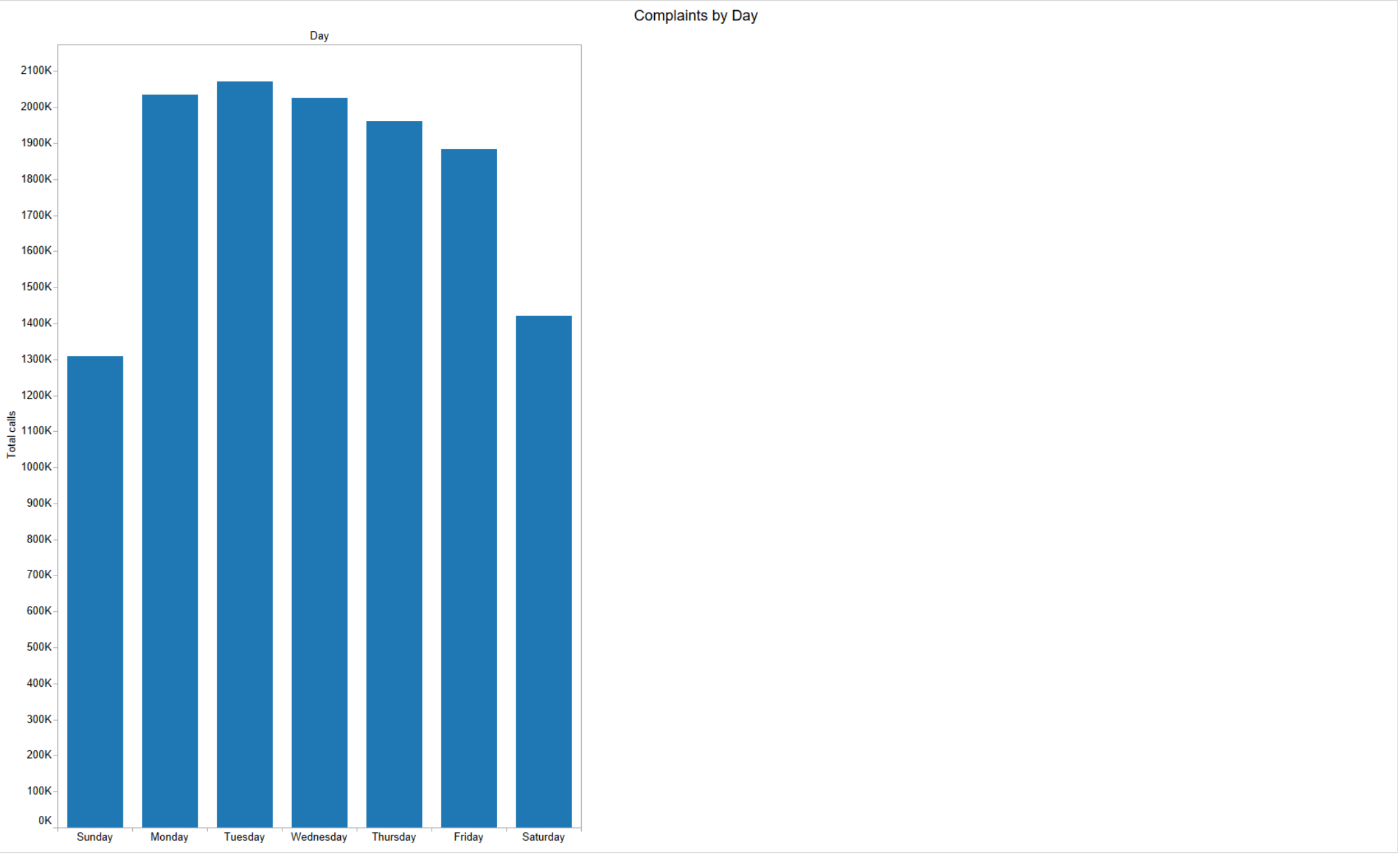
- Heating related complaints were higher at the start and end of the year.
- Street condition complaints were somewhat higher at the start of the year and show a decline after that.
- Water system complaints show a peak in the middle of the year.
- Illegal parking complaints increase slightly in the second half of the year.
- Noise, Street Light Conditions, Blocked Driveway, etc., are some complaints that follow don't vary too much throughout the year.

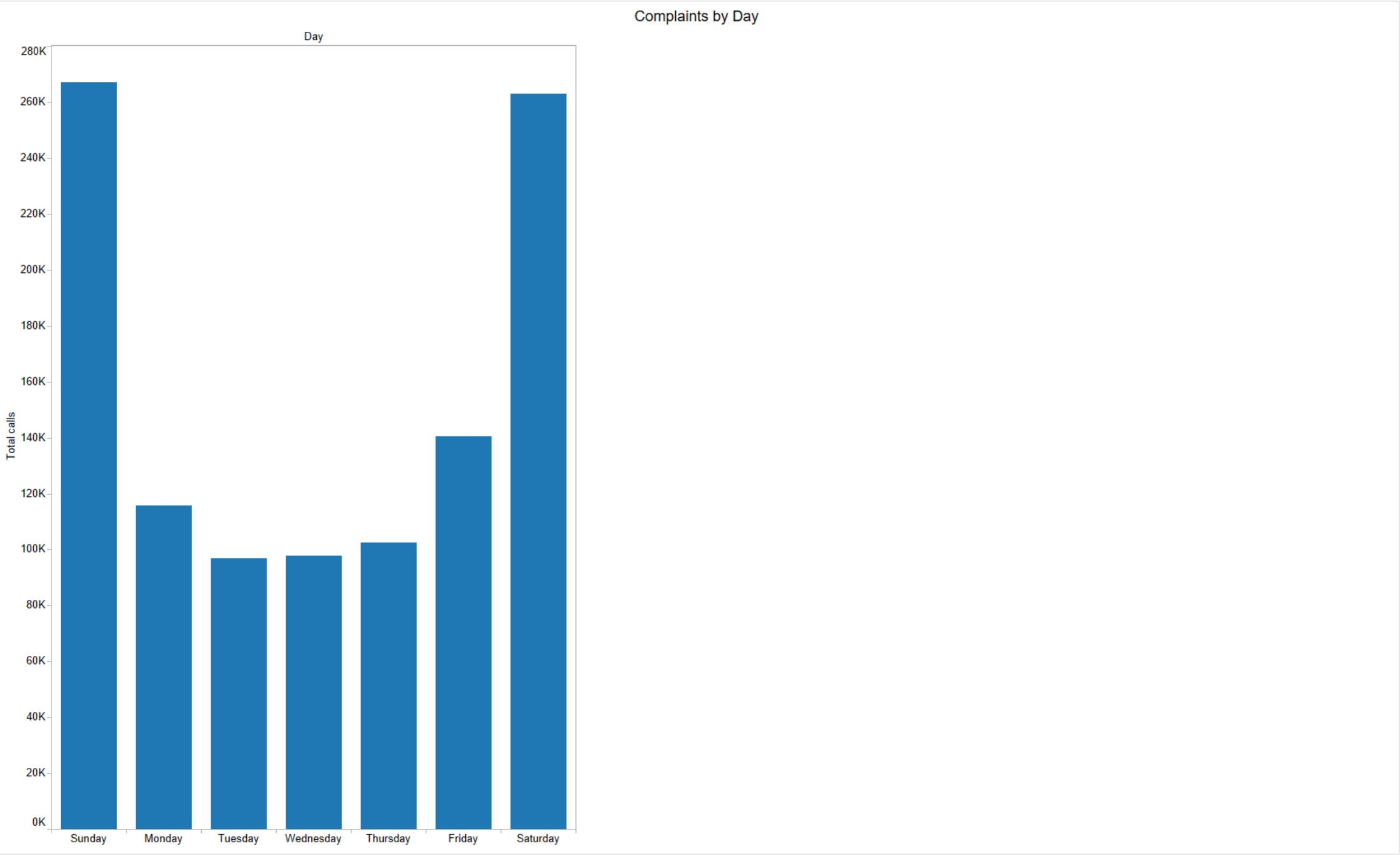
Trends for top 20 Complaints

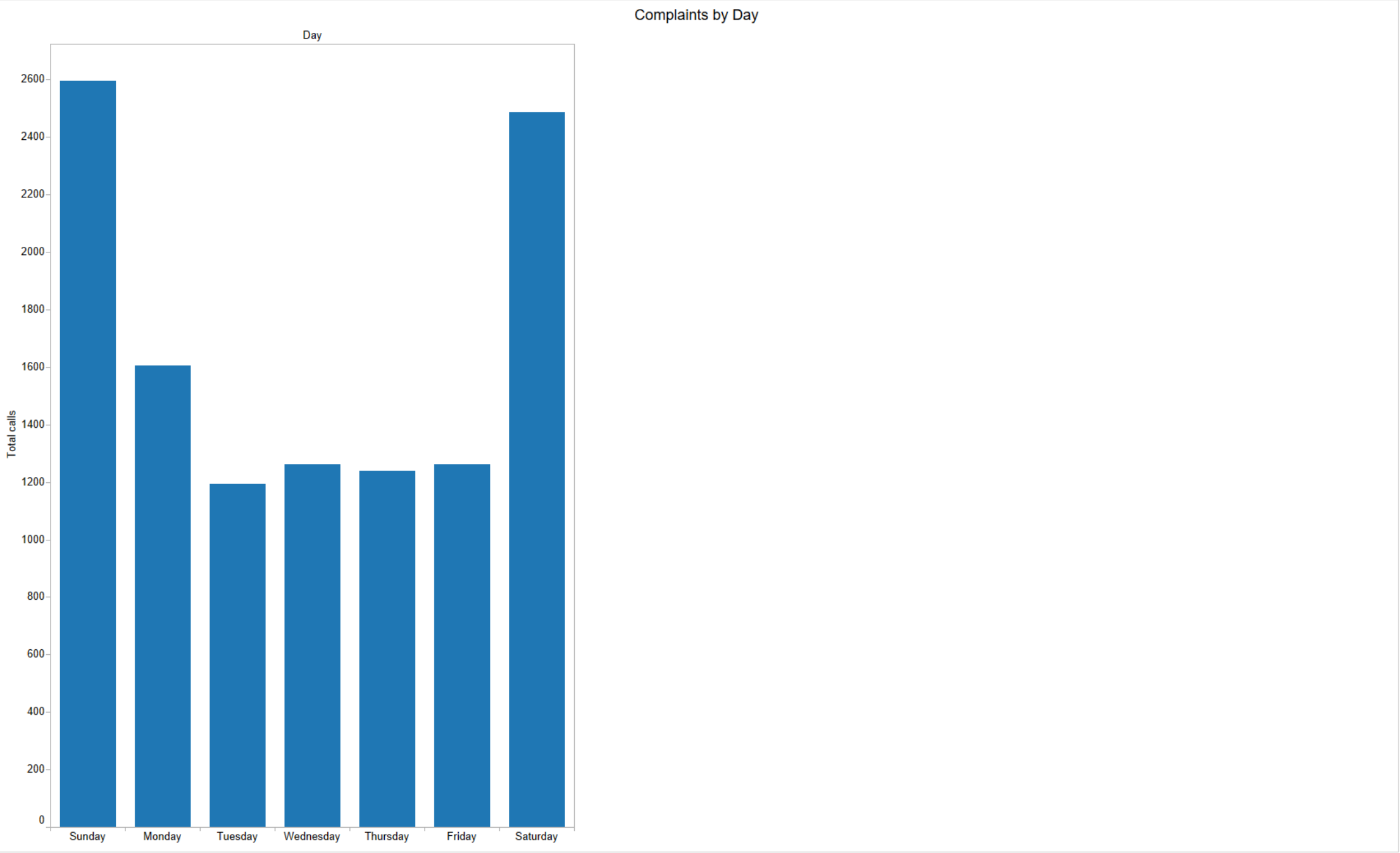


Complaints by day of week

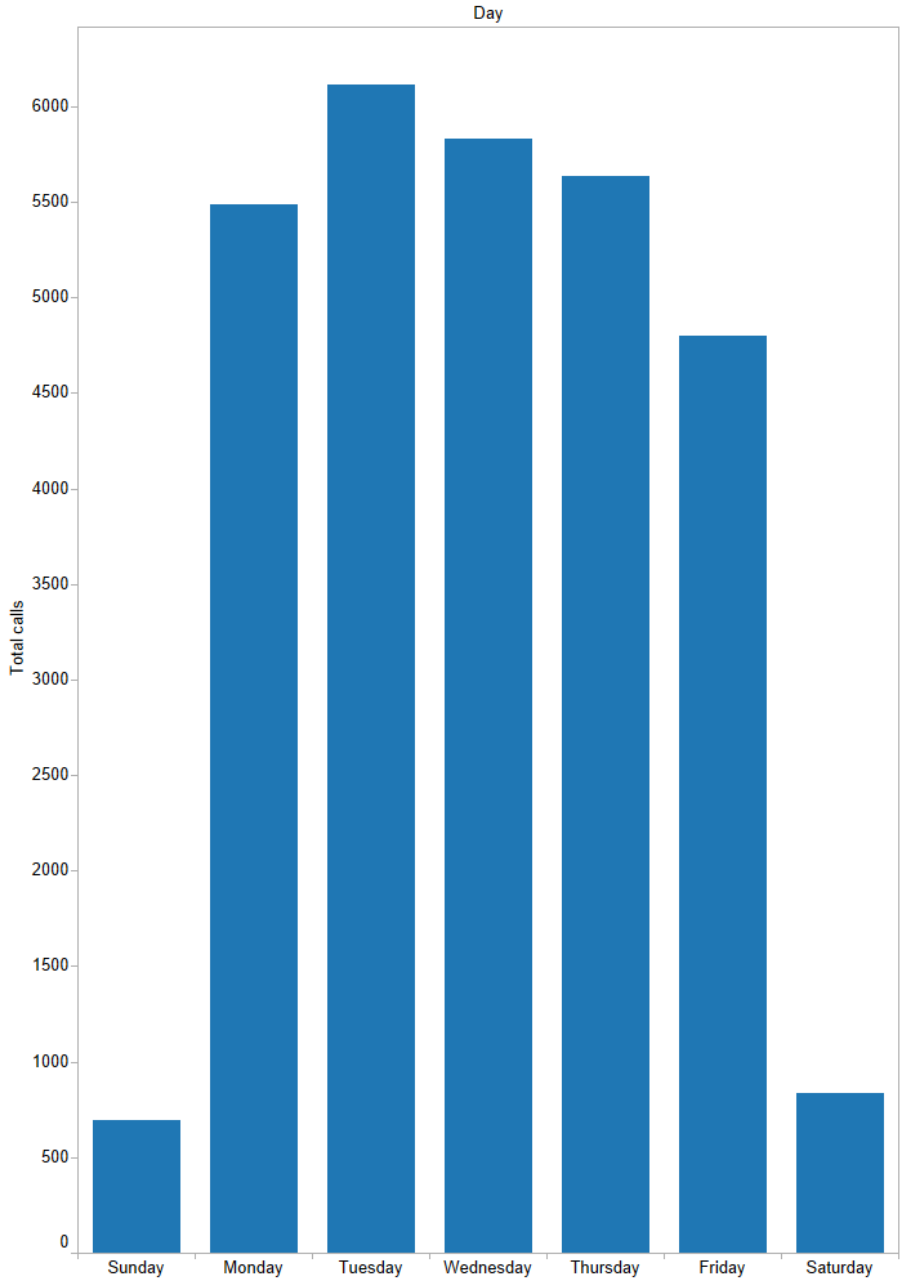
- Complaints were higher in weekdays.
- Noise complaints and Parking Violations were significantly higher in weekends.
- Construction complaints were lower in weekends.
- Blocked driveway complaints were slightly higher in weekends.


Complaints by Day
Day
Total calls
2100K
2000K
1900K
1800K
1700K
1600K
1500K
1400K
1300K
1200K
1100K
1000K
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Complaint
(All)


Complaints by Day
Day
Total calls
280K
260K
240K
220K
200K
180K
160K
140K
120K
100K
80K
60K
40K
20K
0K
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Complaint
Noise - Residential


Complaints by Day
Day
Total calls
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Complaint
Violation of Park Rules

Complaints by Day



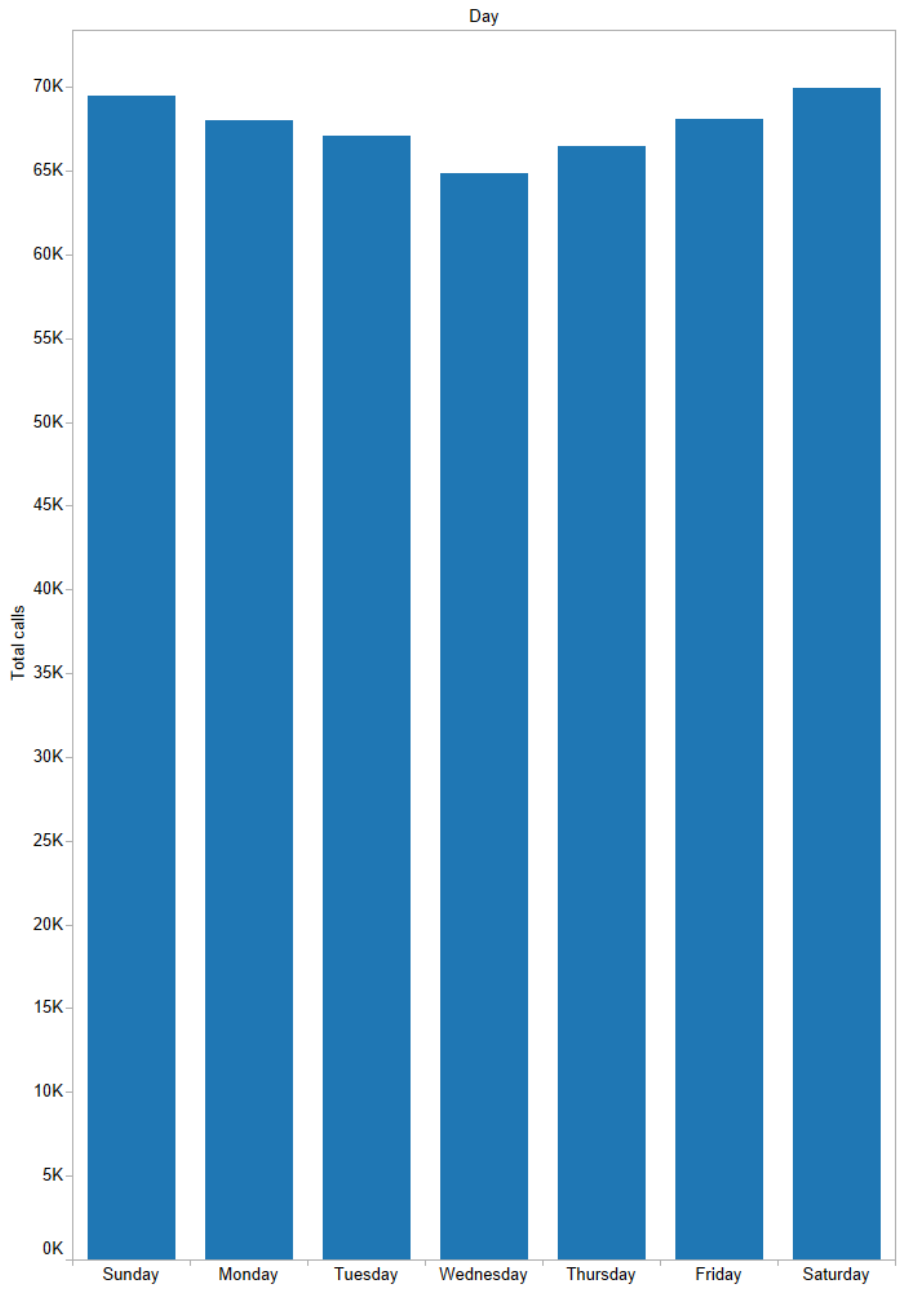
Complaint

(Multiple values)

construction

- [x] CONSTRUCTION
- [x] Construction
- [ ] GENERAL CONSTRUCTION
- [ ] General Construction/Plumbing

Cancel Apply

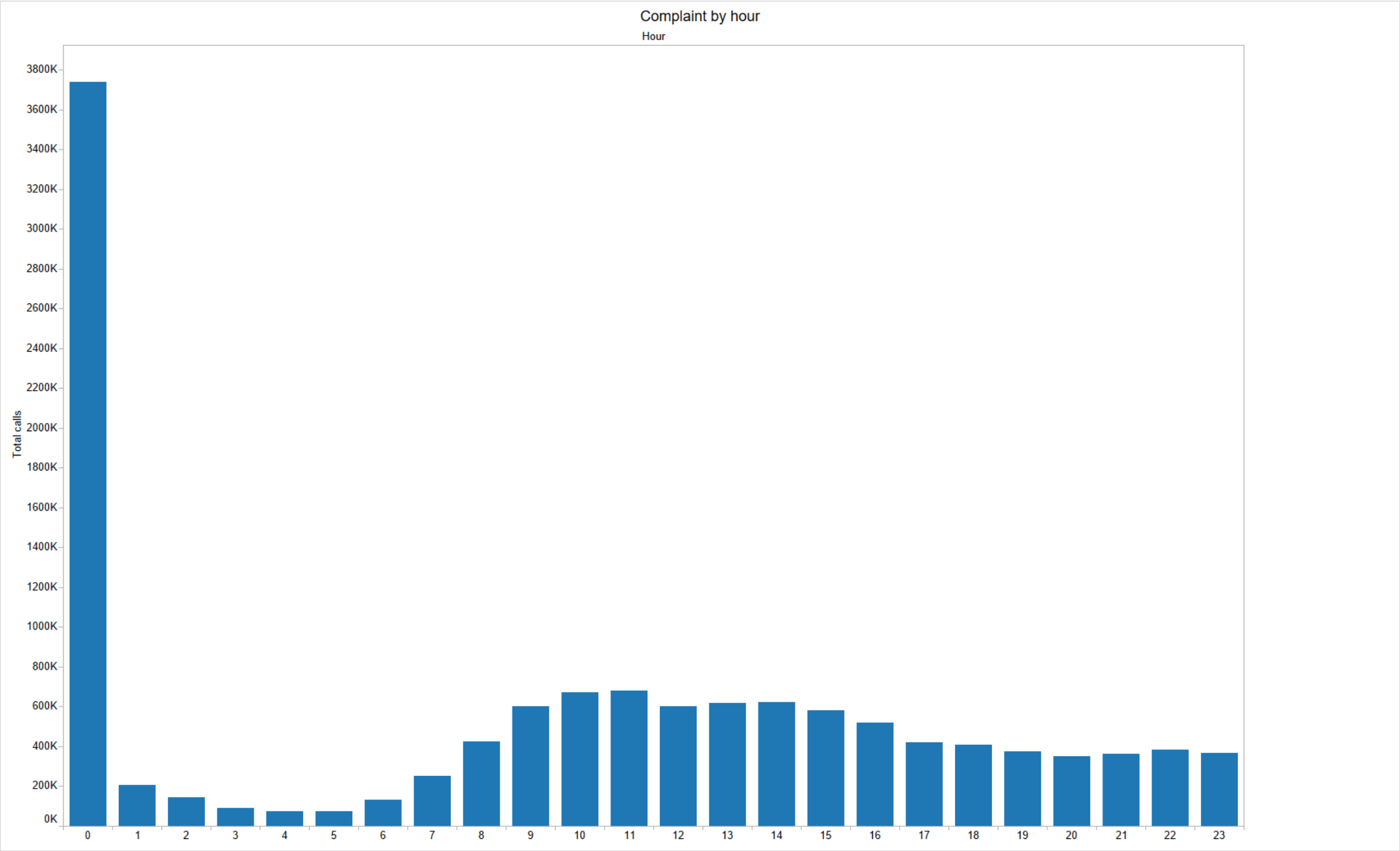


Complaint

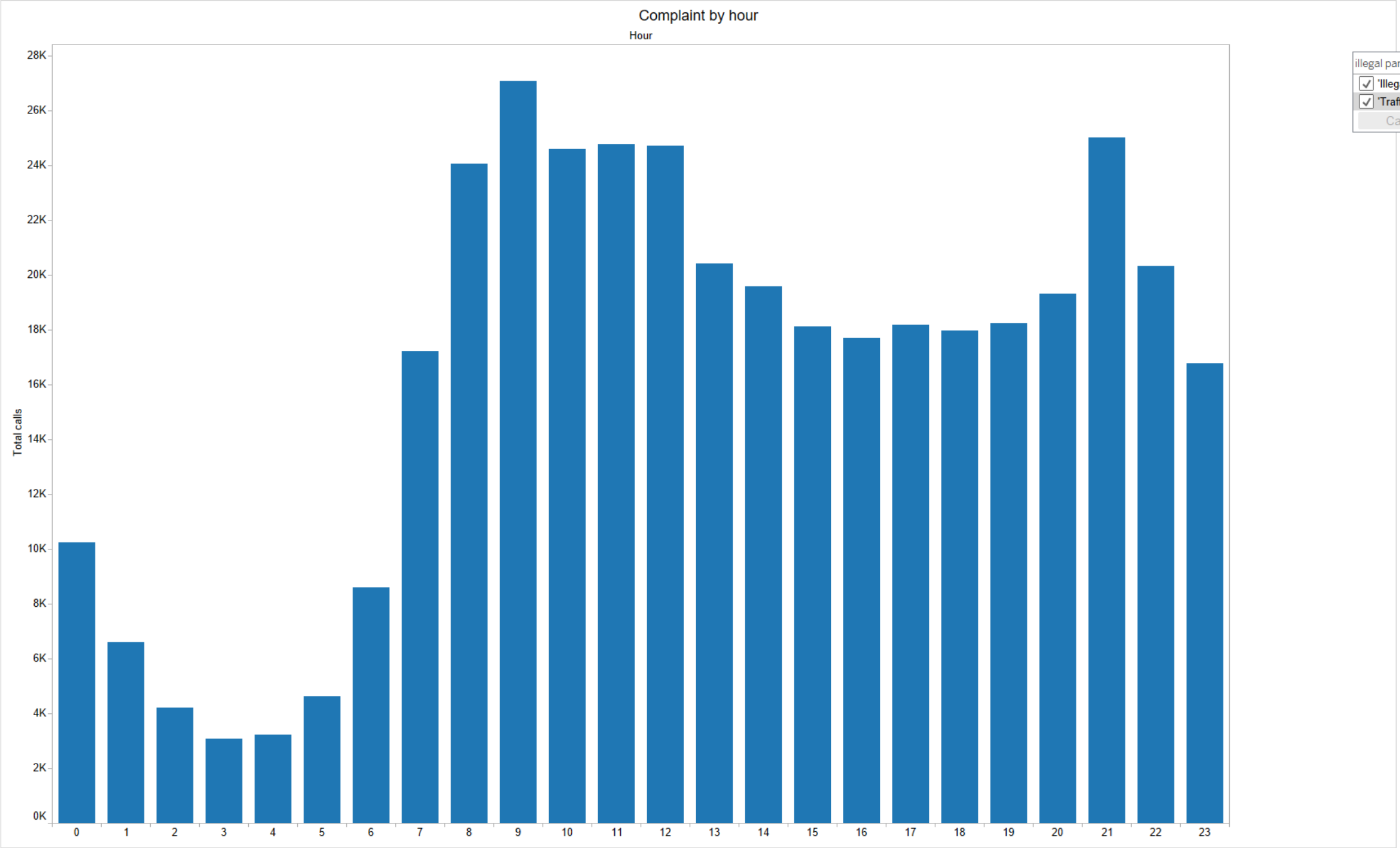
Blocked Driveway

Complaints by time of the day

- Most complaints calls were made close to midnight.
- Illegal parking complaints were high throughout the daytime and fall out after midnight.
- Noise complaints reach a peak close to midnight.
- Parking violation complaints are maximum in the after noon, show a decline after that.
- Blocked driveway complaints are maximum in early morning and night periods.



Map | TreeMap | Complaints TreeMap | Comp Dash | **Complaint by hour** | Average Response time | Complaints by month | Total Complaints



Complaint type

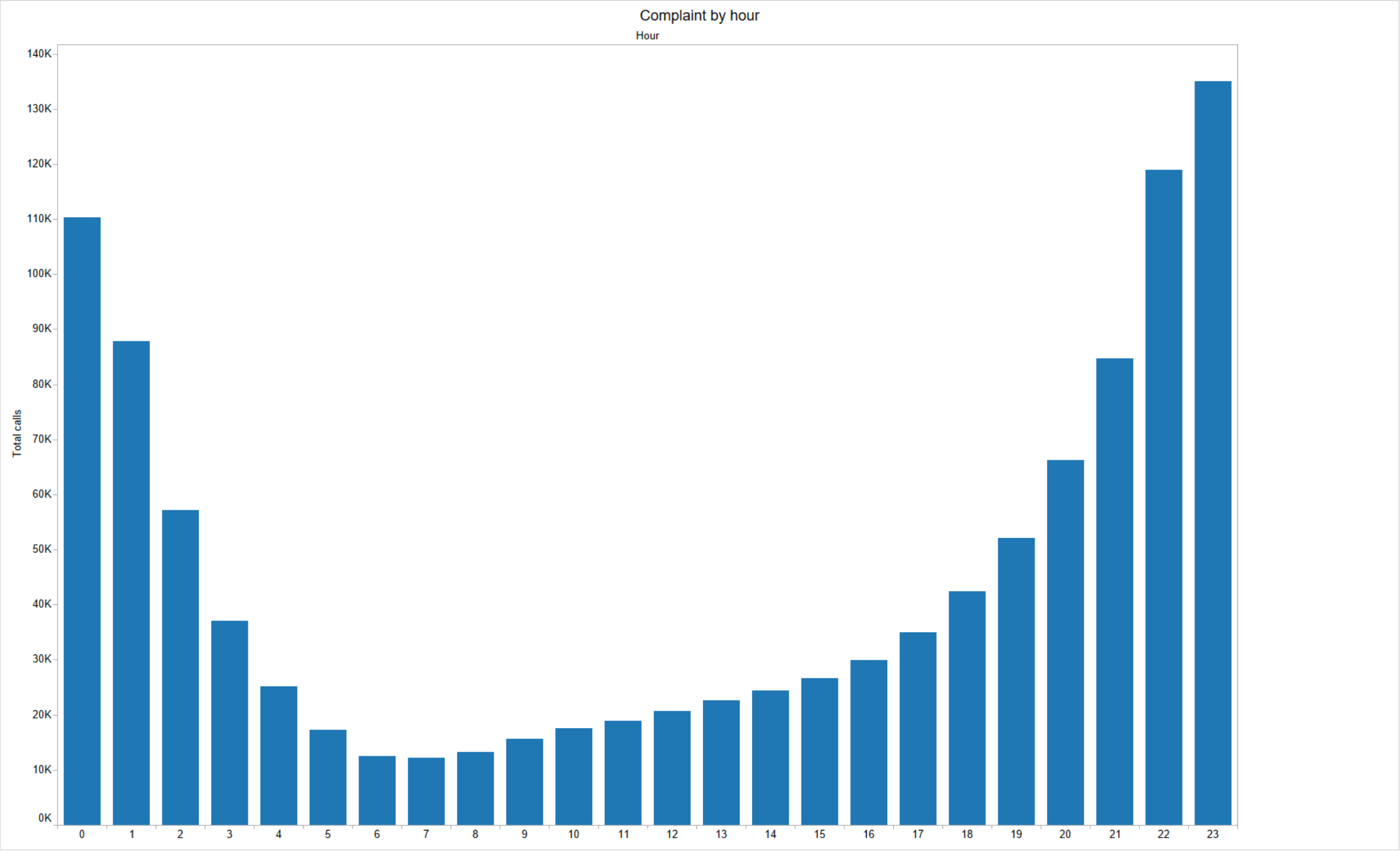
(Multiple values)

illegal parking

- [x] 'Illegal Parking'
- [x] 'Traffic/Illegal Parking'

Cancel Apply

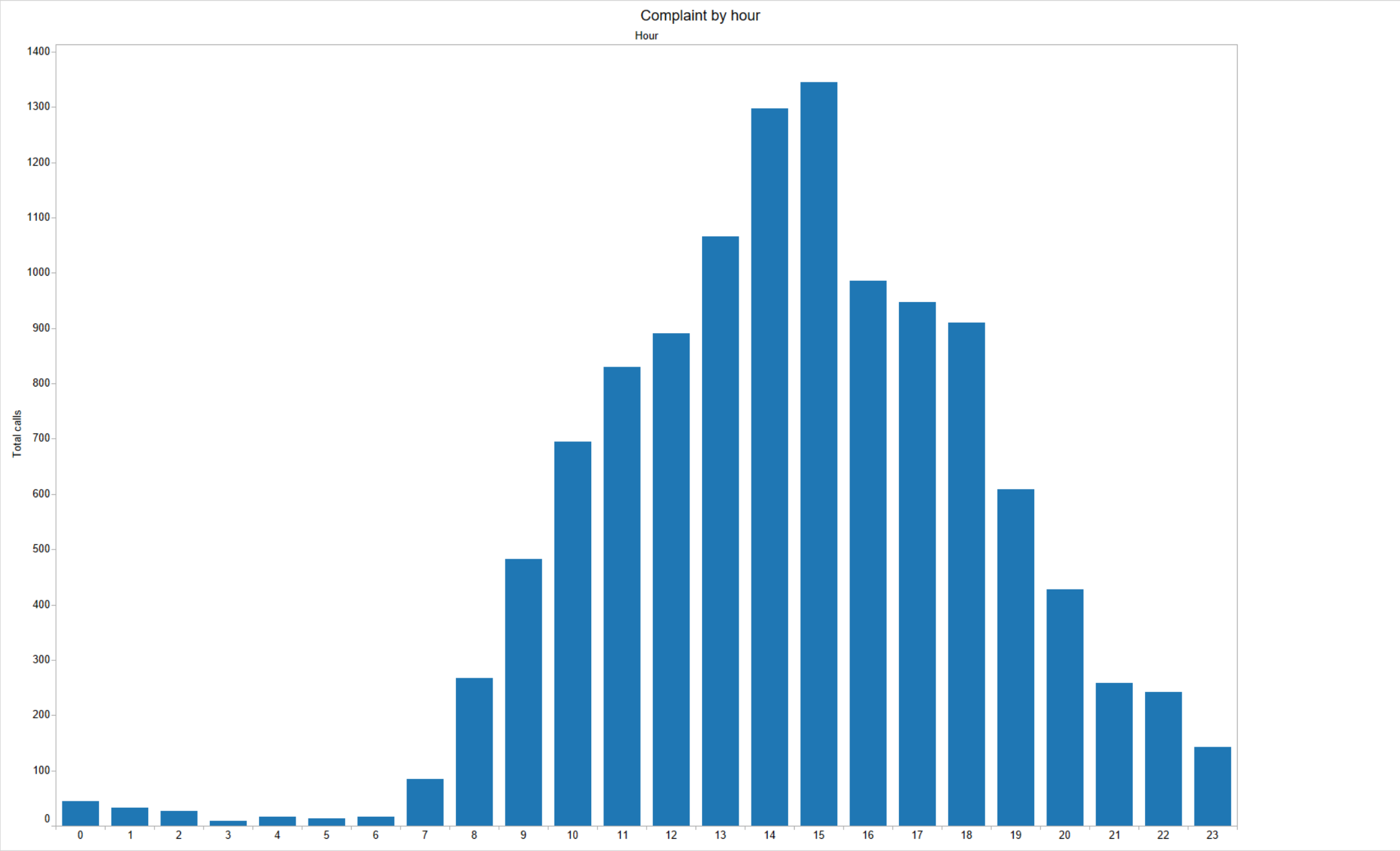
Map | TreeMap | Complaints TreeMap | Comp Dash | **Complaint by hour** | Average Response time | Complaints by month | Total Complaints



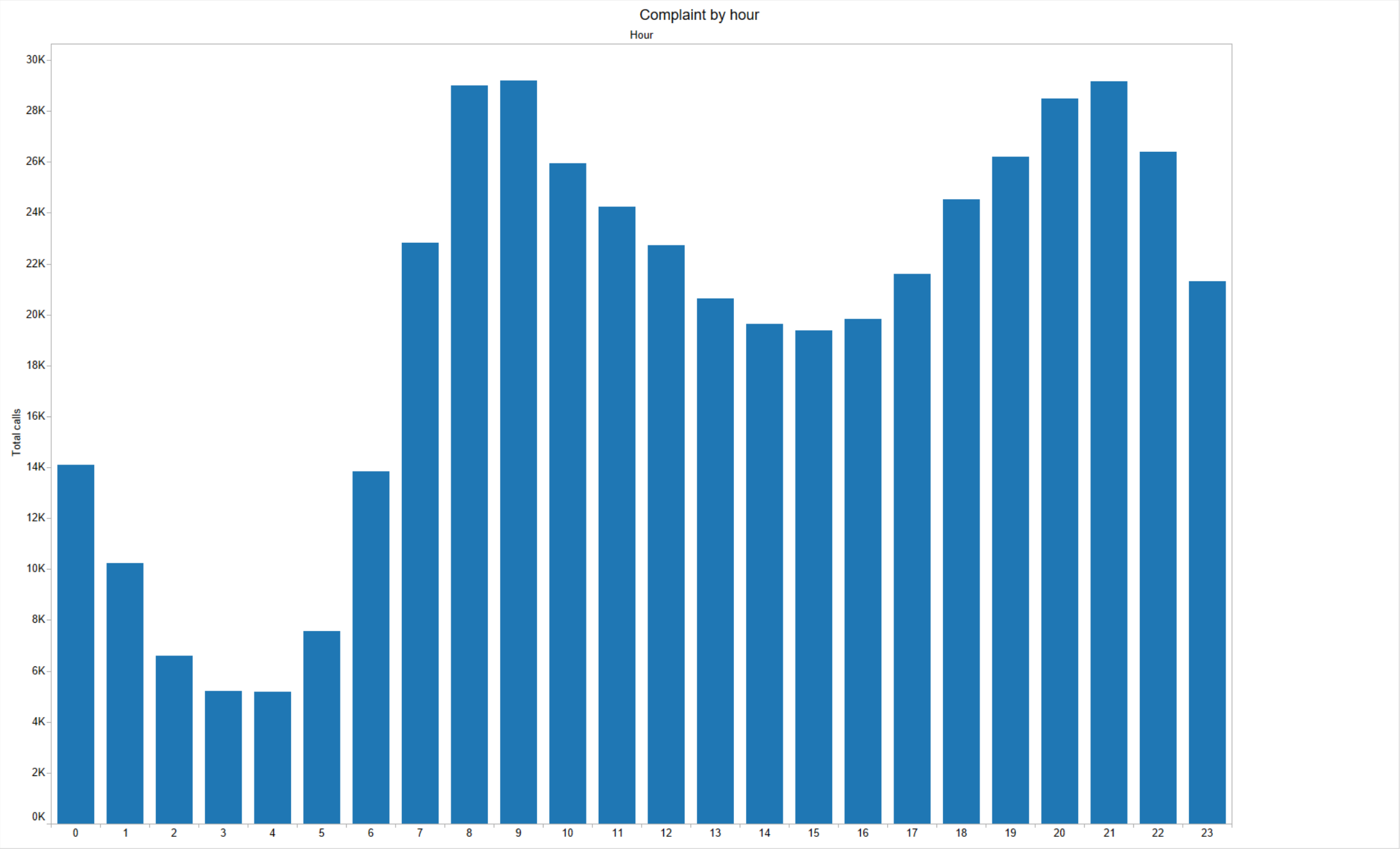
Complaint type

'Noise - Residential'

Map | TreeMap | Complaints TreeMap | Comp Dash | **Complaint by hour** | Average Response time | Complaints by month | Total Complaints



Map | TreeMap | Complaints TreeMap | Comp Dash | **Complaint by hour** | Average Response time | Complaints by month | Total Complaints

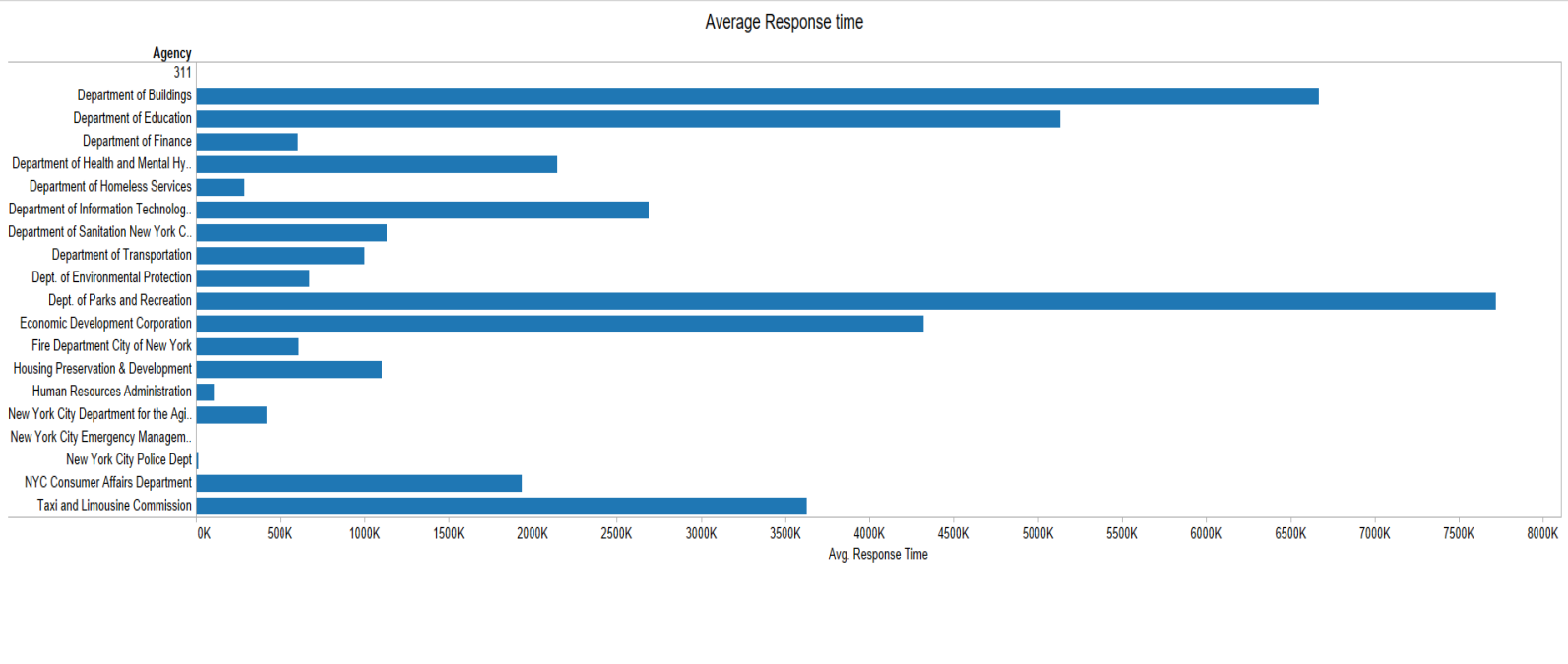


Map | TreeMap | Complaints TreeMap | Comp Dash | **Complaint by hour** | Average Response time | Complaints by month | Total Complaints

Correlating number of calls with average response time

- For each agency the average response time and number of calls was computed using Pyspark.
- Then using the Statistics module the Pearson's R was computed.
- The correlation was weak and negative.
- The Pearson's R was -0.015.
- Hence there is no relationship between the response time and number of calls.

Response times



Predicting response time using random forest

- Response time for the complaints (“Closed Date” – “Created Date”) varies significantly for each agency.
- We use random forest regression algorithm to predict the response time for a complaint.
- The data was given a 70:30 (train : test) split.
- The depth was fixed as 10.
- Features considered were : “Agency”, “Agency Name”, “Complaint Type”, “Descriptor”, “Incident Zip”, “Borough”, “X-coordinate” and “Y-coordinate”.
- The feature “Agency Name” has 1644 different values and hence max bins were 1644.
- The number of trees was set to 3,7 and 16.

Predicting response time using random forest

- The accuracy was computed in terms of the RMSE (root mean square error).
- The results were as follows:
  - 3 trees : 1560 hours (65 days)
  - 7 trees : 1791 hours (74.5 days)
  - 16 trees : 1782 hours (74.3 days).
- 3 trees had the lowest RMSE of 65 days.
- The features considered didn't give very good prediction results.

Conclusion

- Complaint types varied with location, time, day and month.
- Some locations had really high complaint calls.
- Noise was the most reported complaint across all records, followed by heating.
- Response time was not correlated with the number of calls.
- Some agencies have really long response time.
- Random forest regression was used to predict the response time for a complaint.
- However, there was a high error associated with the prediction.

THANK YOU