DESCRIPTION

**Background of Problem Statement:**

NYC 311's mission is to provide the public with quick and easy access to all New York City government services and information while offering the best customer service. Each day, NYC311 receives thousands of requests related to several hundred types of non-emergency services, including noise complaints, plumbing issues, and illegally parked cars. These requests are received by NYC311 and forwarded to the relevant agencies such as the police, buildings, or transportation. The agency responds to the request, addresses it, and then closes it.

**Problem Objective:**

Perform a service request data analysis of New York City 311 calls. You will focus on the data wrangling techniques to understand the pattern in the data and also visualize the major complaint types.  
Domain: Customer Service

**Analysis Tasks to be performed:**

(Perform a service request data analysis of New York City 311 calls)

1. Import a 311 NYC service request.
2. Read or convert the columns ‘Created Date’ and Closed Date’ to datetime datatype and create a new column ‘Request\_Closing\_Time’ as the time elapsed between request creation and request closing. (Hint: Explore the package/module datetime)
3. Provide major insights/patterns that you can offer in a visual format (graphs or tables); at least 4 major conclusions that you can come up with after generic data mining.
4. Order the complaint types based on the average ‘Request\_Closing\_Time’, grouping them for different locations.
5. Perform a statistical test for the following:

Please note: For the below statements you need to state the Null and Alternate and then provide a statistical test to accept or reject the Null Hypothesis along with the corresponding ‘p-value’.

* Whether the average response time across complaint types is similar or not (overall)
* Are the type of complaint or service requested and location related?

============================================================================================================

Finding the missing values from the dataset.

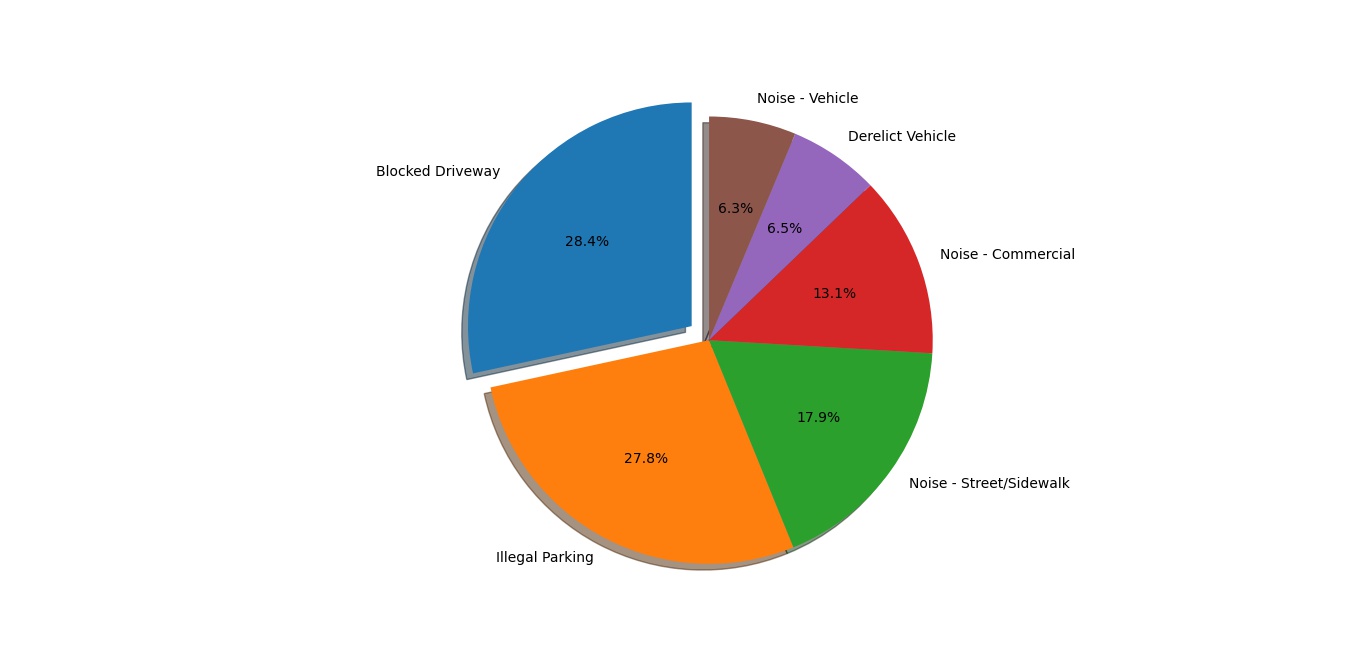
|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Missing Values** | **% of Total Values** |
| School or Citywide Complaint | 300698 | 100 |
| Vehicle Type | 300698 | 100 |
| Taxi Company Borough | 300698 | 100 |
| Taxi Pick Up Location | 300698 | 100 |
| Garage Lot Name | 300698 | 100 |
| Ferry Direction | 300697 | 100 |
| Ferry Terminal Name | 300696 | 100 |
| Bridge Highway Segment | 300485 | 99.9 |
| Road Ramp | 300485 | 99.9 |
| Bridge Highway Direction | 300455 | 99.9 |
| Bridge Highway Name | 300455 | 99.9 |
| Landmark | 300349 | 99.9 |
| Intersection Street 2 | 257336 | 85.6 |
| Intersection Street 1 | 256840 | 85.4 |
| Cross Street 2 | 49779 | 16.6 |
| Cross Street 1 | 49279 | 16.4 |
| Street Name | 44410 | 14.8 |
| Incident Address | 44410 | 14.8 |
| Descriptor | 5914 | 2 |
| Latitude | 3540 | 1.2 |
| Longitude | 3540 | 1.2 |
| Y Coordinate (State Plane) | 3540 | 1.2 |
| Location | 3540 | 1.2 |
| X Coordinate (State Plane) | 3540 | 1.2 |
| Address Type | 2815 | 0.9 |
| Incident Zip | 2615 | 0.9 |
| City | 2614 | 0.9 |
| Resolution Action Updated Date | 2187 | 0.7 |
| Facility Type | 2171 | 0.7 |
| Closed Date | 2164 | 0.7 |
| Location Type | 131 | 0 |
| Due Date | 3 | 0 |
| School Zip | 1 | 0 |
| School Code | 1 | 0 |
| School Region | 1 | 0 |

Observations:

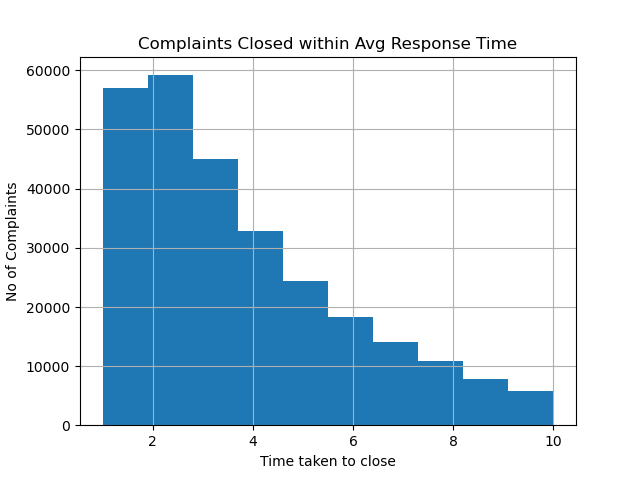
1. Post checking the missing value % it was found that out of 53 features/columns, 14 columns had almost 100% missing values or NaN. Hence the same were dropped from the dataset.
2. The highest no of complaint type is ‘Blocked Driveway’ which pertains to 26% of the total complaints followed by Illegal Parking (25%), Noise – Street/Sidewalk (16%), Noise – Commercial (12%), Derelict Vehicle – (6%) and Noise Vehicle (6%)

|  |  |  |
| --- | --- | --- |
| **Complaint Type** | **Count** | **%Contri** |
| Blocked Driveway | 77044 | 26% |
| Illegal Parking | 75361 | 25% |
| Noise - Street/Sidewalk | 48612 | 16% |
| Noise - Commercial | 35577 | 12% |
| Derelict Vehicle | 17718 | 6% |
| Noise - Vehicle | 17083 | 6% |

**Presenting graphically:**



1. Maximum time taken to resolve a complaint was for 24 days and the minimum time taken is 0:01:00 minute.
2. Average time taken to resolve a complaint is 04:18:51 (HH:MM:SS).



**Hypothesis Testing:**

Statement: - Whether the average response time across complaint types is similar or not (overall)

H0 = Average response time for all complaints is same.

H1= Average response time for at least one complaint type is different.

1. **NORMALITY TEST for Request\_Closing\_hours** – We perform Shapiro Wilk’s test to check the normal distribution of a sample. If P-value is high then we fail to reject Null Hypothesis else we reject Null Hypothesis

H0:- Here we assume that the distribution is normal.

H1:- Distribution is not normal.

After performing the test it was observed that the P-Value for Request\_Closing\_hours is 0.51 which is > 0.05. Hence we fail to reject Null Hyp.

Similarly for Complaint Types – the p value was 0.602 which is > 0.05 hence we fail to reject Null Hyp.

1. **ANOVA Test - One way ANOVA** is performed to The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups (although you tend to only see it used when there are a minimum of three, rather than two groups).

After performing the f\_oneway test we observed that the p-value =0.0 hence we will be rejecting the H0 which is Avg response time for all complaints are same.

1. Chi Square Test – This test is performed when we want to compare between two categorical values if they are related or not.

H0 = Location and complaint type are related to each other.

H1= Both location and complaint type are non-relative to each other.

After performing the ChiSquare contingency test we observed that the p value is 0.00 which is less than 0.05 hence we reject Null Hyp.