My Conclusions after looking into file “pager.xlsx”

**What are the top features that effect the time taken?**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Agency | Important |
| Complaint Type | There is spread in the categories |
| Descriptor | Needs preprocessing of text and combining of categories to reduce the number of categories |
| Location Type | Useful but needs little preprocessing of the text in the columns |
| Borough | with little amount of preprocessing, drop the unwanted records |

I don’t mean to say that other are not important but they need some changes before they can be used.

Ex: City is also important but it needs to be preprocessed because there typos in the values

“adena” “adina” mean the same

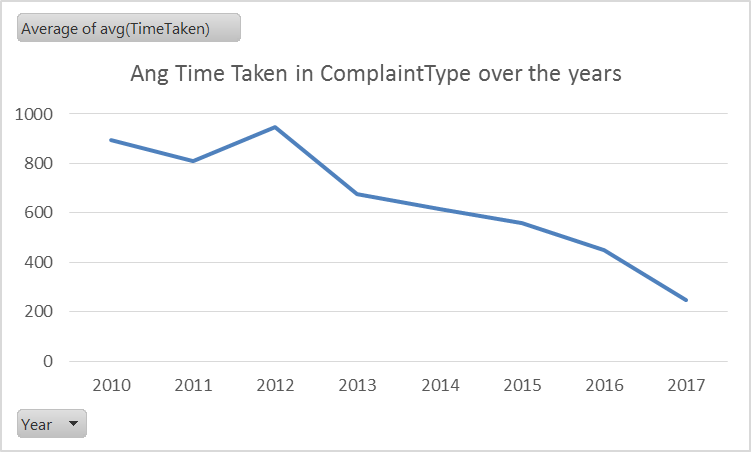
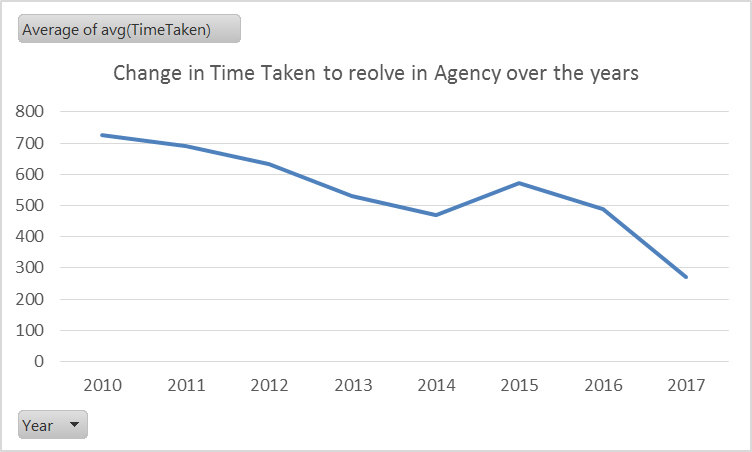
Please see **column\_analysis\_description.xlsx** for more details

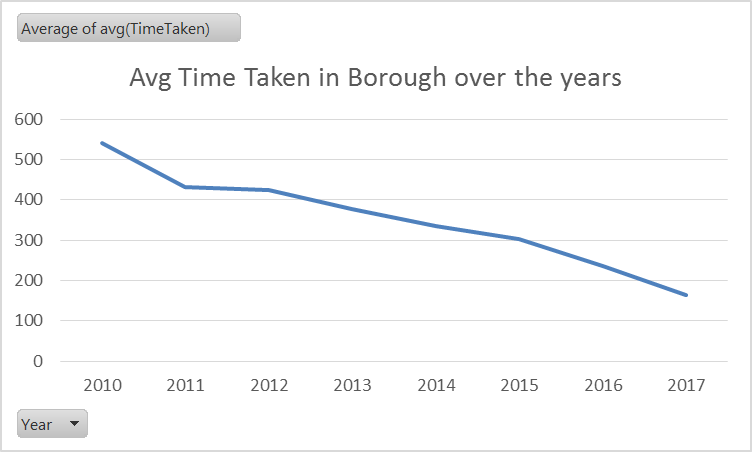
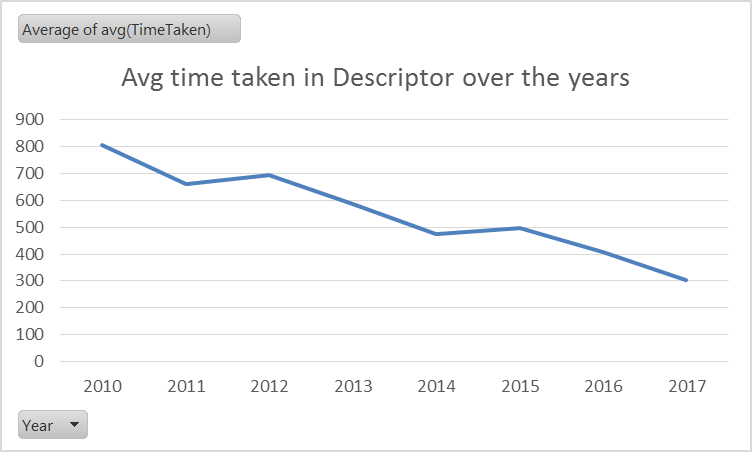
**Do the top features they change over time?**

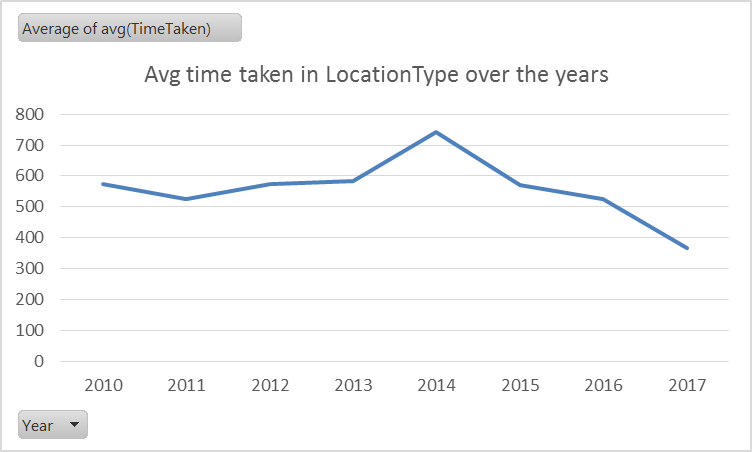
Yes, the time taken **reduces** across **years** but across **months** much **no** change is noticed.

Please see **pager\_yearly\_analysis\_description.xlsx** for more details

|  |  |
| --- | --- |
| **Column** | **Change** |
| Agency | Decreases |
| Complaint Type | Decreases |
| Descriptor | Decreases |
| Location Type | NA |
| Borough | Decreases |



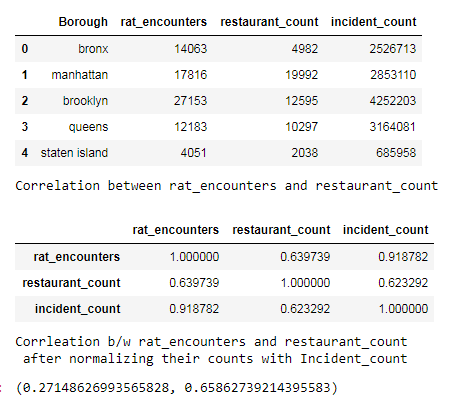




Please see **pager\_montly\_analysis\_description.xlsx** for more details

* No changes noticed for graphs please look into the excel file

**Are rat sightings more common in areas with higher density of restaurant’s with low inspection grades?**



So from the correlation we can say rat encounters are not related to restaurant count with low inspection grades.

In the first table there seems to be co-relation but that is due to large number of incidents being reported in that particular regions

**Two Additional Questions**

1. Are the NYPD police officers/NYPD offices located to close to the areas from where more incidents are reported?

To answer this question we need to know the locations of the officer’s. We can run GMM algorithm on NYPD incident data to get the centers from where more incidents are reported and GMM algorithm on NYPD officer’s locations to check if the officers are located at the right location to reach the incident zone in optimal time.

We can then manipulate the locations to improve response times.

1. Do more people fall sick in regions with higher number of incidents?

To answer this question we need to have healthcare sickness/diseases data/health-insurance data.

**Dive Deep**

We can study if less sanitation is a primary reason for people falling sick. The take steps to improve sanitation in that particular areas and then check if improving sanitation has contributed to improvement in health of the people. This can be a complicated study as we have to consider several factors into account and then take the large enough sample to support the study.

We need to do significance testing in which we have to choose a confidence level and then be able to prove the significance of claim.

Similarly, we can also study if crime is related to health of the people.

Several factors need to be taken into consideration when designing such tests for which a data scientist need to sit with the domain experts to understand the domain before making assumptions or conclusions.