Water Quality Complaints

***Introduction***

New York City is known for its high-quality tap water that is safe and delicious to drink. However, the DEP (Department of Environmental Protection) still receives over 800 complaints per year from New York residents regarding their water quality. I pulled a dataset from NYC OpenData regarding these complaints[[1]](#footnote-1). This data shows complaints placed via 311 from 2010 until mid-2018. The data includes information on when and where the complaint was made, along with a description of the complaint and its resolution status.

The data I pulled had information regarding complaints ranging from 2010 up until July of 2018. However, the analyses I ran all excluded 2018 due to its incomplete data. While these analyses could have included 2018 data, this would have required the use of data analysis/mining techniques to predict or model the rest of the year. Instead, I decided to focus on the data we have available to evaluate the agency’s possible inefficiencies.

Though the dataset is small enough that it can be easily manipulated in Excel using PivotTables and formulas, I also wrote code using Python Pandas to provide scalability to the project. As more data is collected, running Python code for an analysis becomes quicker and more efficient than using Excel. All graphs included in the report were made using the chart capabilities in Excel.

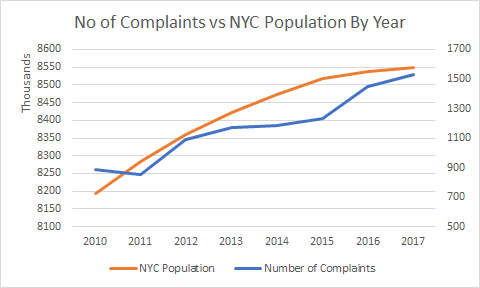
***Analyses***

Looking through the dataset, I wanted to answer some preliminary questions and then dive further into questions regarding possible agency inefficiencies:

1. How is the number of complaints changing over time?

I ran my code [1] to observe how the number of complaints has been trending since 2010. The results can be found in the Python Code Appendix. Despite the small dip in 2011, there is a general upward trend in the number of complaints per year. I decided to graph the number of complaints against the population growth in NYC to gain more context into the rise of water quality complaints. Figure 1 displays the two trends and both are moving upwards; NYC’s population growth has been relatively smooth and regular while the increase in complaints seems to be less predictable. While a more in-depth analysis would be needed to determine whether population growth is the sole or dominant factor driving the increase in complaints, the fact that both quantities have consistently increased over the timeframe in question is not an immediate cause for concern. Had the population decreased or stayed the same, and the complaint volume still increased, this would be a sign of possible decreasing DEP effectiveness.

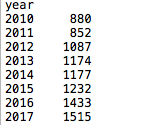
*Figure 1: No. of Complaints Compared to NYC Population Growth by Year*



2. Is the agency getting better at closing complaints?

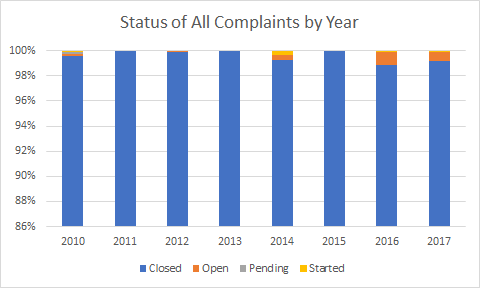
To answer this question, I investigated the number of closed complaints per year to see how well the DEP is doing in terms of their resolution rate. You can find the code for this in the Python Code Appendix [2]. Figure 2 shows the Python output of the number of closed complaints per year. Though there is a dip in closed complaints from 2010 to 2011, the general trend is upward in the number of resolved complaints per year.

*Figure 2: Closed Complaints per Year*



To put this trend into context, I calculated the ratio of closed to open/pending complaints by year. Figure 3 displays the breakdown of all complaints. As you can see, closed complaints make up at least 98% of all complaints received. The year 2016 saw the most open complaints, but that still only made up 1% of all complaints that year, which is negligible. Both the years 2011 and 2013 had 100% rate of complaints closed by DEP.

*Figure 3: Status of All Complaints by Year*

**

Investigating further, I wanted to see which zip codes had the most open and pending complaints [3].

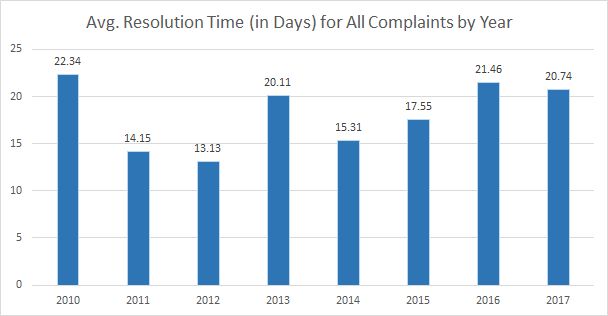
After running the code, I saw that the zip codes with the highest number of open and pending complaints since 2010-2017 have been 10471, 10025, 11421, 11363, 10469, and 11217. However, each of these zip codes still only has two open or pending complaints, which is not a statistically significant difference relative to the majority of zip codes with no open complaints. Therefore, this analysis does not indicate any structural inefficiencies at the DEP.

3. What is the agency’s average resolution time per year? Which zip codes have the longest resolution times?

Though it appears that most complaints eventually get closed, it would be helpful to see how long it takes for complaints to get resolved. To do this, I ran my code [4] to find the average resolution time for all complaints per year. I subtracted the Closed Date from the Created Date and filtered any values that didn’t make sense (negative time values where the Closed Date came before the Created Date).

Figure 4 shows the average amount of time, in days, it has taken over the years to resolve all complaints. The year 2012 had the most efficient average time, approximately 13 days, and resolution time has gone up significantly in the last three years, with 2017 having an average resolution time of almost 21 days per complaint. As we saw earlier, the amount of complaints has been increasing since 2011, which could be a reason behind the increase in resolution times. Looking further into DEP funding and how many workers are currently allocated to handle complaints would give us a better idea of whether misuse is taking place.

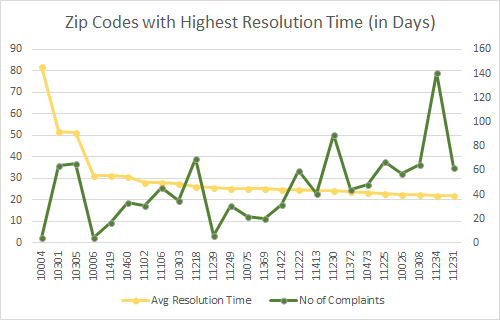
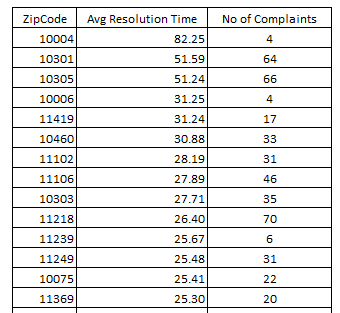
*Figure 4: Avg. Resolution Time for All Complaints by Year*



Digging further into this, I was curious to see the breakdown of the resolution times per zip code and if there are any areas that have longer resolution times than others [5]. After running my code, I saw that some zip codes had resolution times of up to 82 days, which is almost 3 months. However, to avoid looking at data that might be skewed because it’s taking the mean, I graphed the top 25 zip codes with the longest resolution times against how many complaints they’ve filed in all. It’s easy for an average to be skewed when there aren’t enough data points so outliers can easily influence the final calculation. Plus, intuitively, you’d expect the two lines to track each other since the influx of complaints would be a reasonable indicator into why response time might be slower for a particular zip code.

Figure 5 displays this graph where the yellow lines shows the average number of days a zip code’s complaints take to get resolved while the green shows the number of complaints that zip code has filed. Despite the first zip code having a turnaround time of 82 days, it only has 4 complaints so would need further investigation. The next two zip codes are better suited since they each have a significant number of complaints to point to a trend. Both zip codes 10301 and 10305 are in Staten Island. According to my code [6], the average resolution time across all zip codes is about 18 days. There are several zip codes here that could point to misuse as their average resolution time is notably higher than the overall average for all zip codes, even taking the yearly breakdowns into consideration. Is there a problem in providing service to these areas for some reason?

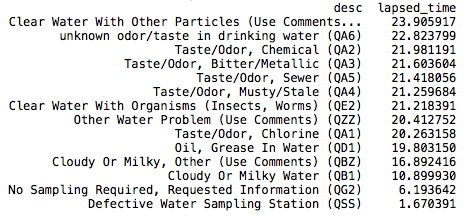
*Figure 5: Zip Codes with Highest Resolution Time*

* *

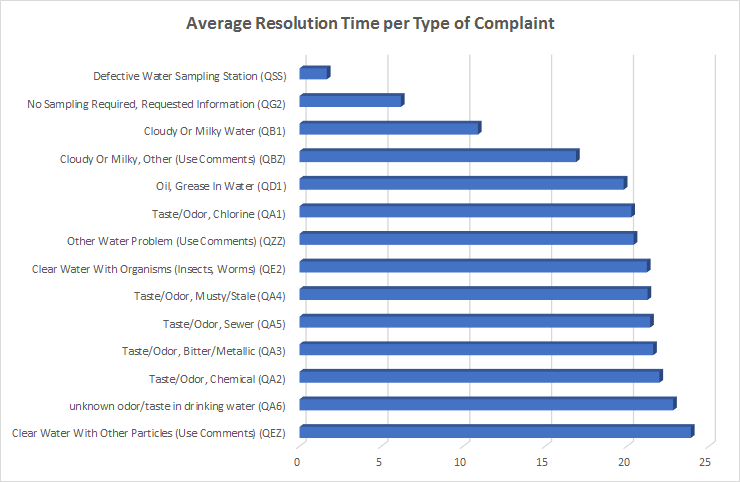
As I pointed out earlier, these two lines should roughly track each other assuming that the volume of complaints is the main driver of response time in given neighborhood. However, looking at the data, this does not appear to be the case. The resolution times hangs at around 21-24 days despite the volume of complaints coming in for a particular zip code. Based on this limited data, I would direct efforts to improve resolution time to look at factors other than complaint volume per neighborhood.

To follow through with my previous thought, I considered that maybe these zip codes had such high resolution times due to specific water quality complaints taking longer than others. I ran my code [7] to see the average resolution time per type of water complaint. Figure 6 displays the Python results while Figure 7 graphs them. I saw that most water quality complaints took about 19-23 days to resolve despite its type. The only outliers are the complaints QSS, QG2, QB1, and QBZ. Both QSS and QG2 are the only complaints that deal with water sampling issues so I’m assuming those are quick to get resolved. QB1 and QBZ are complaints regarding Cloudy or Milky Water. This is also the most common complaint, making about 16% of all complaints placed by all boroughs. I’m assuming this complaint is also resolved quickly since it’s so common[[2]](#footnote-2) and is a harmless phenomenon that often occurs during the colder months[[3]](#footnote-3). Concluding from this analysis, complaint type should not have that much of an impact on these zip codes’ high resolution time so there must be other factors at play.

*Figure 6: Average Resolution Time per Complaint Type*



*Figure 7: Average Resolution Time per Complaint Type*



Conclusions

* The number of water quality complaints has been on the rise since 2010, however, this could be attributed to NYC population growth or the awareness of global environmental concerns.
* It would appear from findings, that there is a high resolution rate in terms of the amount of complaints that are closed amongst the DEP’s water quality complaint complaints, which is a sign of efficiency. Though some zip codes have higher number of open and pending complaints than others, the number of these is statistically insignificant compared to those that are closed.
* It looks like the resolution time for all complaints has been increasing in the last couple of years. Looking further into DEP funding and how many workers are currently allocated to handle complaints would give us a better idea into the possible misuse taking place. Furthermore, some zip codes have a notably higher average resolution time than others, despite complaint volume and complaint types. This should be investigated further.

Further analysis could include supplementing the data with zip code population to see the ratio in the number of complaints per zip code to its population. Are there any zip codes where the population is small compared to how many complaints are placed? This could also point to agency problem areas.

**PYTHON CODE APPENDIX**

Note: all code starts with the following

import pandas as pd

import datetime

df = pd.read\_csv("water\_complaints.csv")

df = df[["Created Date","Closed Date", "Status", "Unique Key","Incident Zip","Descriptor"]] *#selecting only the columns I need*

df.columns = ["created", "closed", "status", "comp", "zipcode", "desc"] *#renaming columns for easier use*

df["created"] = pd.to\_datetime(df["created"])

df["closed"] = pd.to\_datetime(df["closed"])

df = df[(df["created"]<datetime.date(2018,1,1))] *#filtering out any 2018 complaints from my data*

df["year"] = pd.DatetimeIndex(df["created"]).year *#creating a new column for year a complaint was placed*

df["lapsed\_time"] = (df["closed"] - df["created"]).dt.days *#creating a new column for resolution time*

1. Calculating total number of complaints by year

gf = df.groupby(["year"])

gf = gf.count()

print(gf["comp"])

2.  Calculating total number of closed complaints per year and then the percentage of closed complaints by year

gf = df.loc[df["status"] == "Closed"]

gf = gf.groupby(["year"])

gf = gf.count()

print(gf["comp"])

--------------------------------------------------------------

tf = df.groupby(["year"])

tf = tf.count()

total = tf["comp"]

gf = df.loc[df["status"] == "Closed"]

gf = gf.groupby(["year"])

gf = gf.count()

closed = gf["comp"]

perc = closed/total

perc = perc\*100

perc = perc.round(2)

print(perc)

3. Finding zipcodes with the highest number of open and pending complaints

gf = df.loc[df["status"].isin(["Open", "Pending"])]

gf = gf.groupby(["zipcode"])

gf = gf.count()

gf = gf.sort\_values("comp", ascending=False)

print(gf["comp"])

4. Calculating average resolution time per year

gf = df[(df['lapsed\_time']>=0)] *#filtering out any resolution times that don’t make sense like negative values*

gf = gf.groupby(["year"], as\_index=False).mean()

print(gf[["year","lapsed\_time"]])

5. Calculating average resolution time per zipcode

gf = df[(df['lapsed\_time']>=0)]

gf = gf.groupby(["zipcode"], as\_index=False).mean()

gf = gf.sort\_values("lapsed\_time", ascending=False)

print(gf[["zipcode","lapsed\_time"]])

6. Calculating the average resolution time across all zipcodes

gf = df[(df['lapsed\_time']>=0)]

gf = gf.mean()

print(gf["lapsed\_time"])

7. Calculating the average resolution time per each complaint type

gf = df[(df["lapsed\_time"]>=0)]

gf = gf.groupby(["desc"], as\_index=False).mean()

gf = gf.sort\_values("lapsed\_time", ascending=False)

print(gf[["desc","lapsed\_time"]])

1. <https://data.cityofnewyork.us/Environment/Water-Quality-complaints/qfe3-6dkn/data> [↑](#footnote-ref-1)
2. <https://water.usgs.gov/edu/qa-chemical-cloudy.html> [↑](#footnote-ref-2)
3. <http://www.nyc.gov/html/dep/pdf/wsstate16.pdf> - page 23 [↑](#footnote-ref-3)