Daniel Aguilar-Rodriguez
Jia Cong Lin
Anvinh Truong

12/16/2022
Introduction to Data Science

BLOG: Air Quality in NYC

Motivation

For our group, what we are trying to solve and predict is the air quality in New York City based on the traffic levels. It's important to take into consideration the air quality and what we breathe in, especially in a dense place like New York City. There will be times when the air quality will be unhealthy to breathe in and there will also be times when it is fine, so it would be helpful to predict future air quality based on the traffic volumes and determine whether or not the air quality would be safe to inhale on any given day.

Data

For the datasets, we have one on Air Quality and another on Traffic Volume, with the Air Quality data being sourced from the EPA, and Traffic Volume from NYC OpenData. The Air Quality dataset contains the type of sample in the air, how it's measured, neighborhood, time frame, and start date, which are all relevant in having a more precise measurement in a particular area in New York City. The Traffic Volume dataset contains the boroughs for where the total traffic is measured within 15 minute increments, the streets from the start to end at where the traffic is located, and date/time at when it took place. These two datasets would help us predict air quality based on the amount of traffic there is, however there are some limitations to this in which we are not taking the weather into account as that can have an impact on both air quality and traffic.

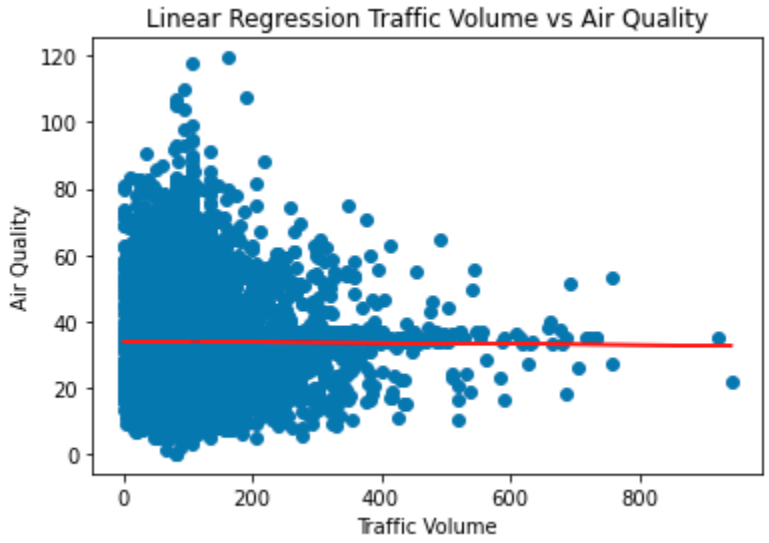
For the Air Quality dataset, some limitations in it are with how the time frame is within seasons and years, and it couldn't really be used with the Traffic Volume dataset because of how a more precise date couldn't be found. It was harder to consolidate a date that would work with both the Air Quality and Traffic Volume datasets. In the Air Quality dataset from the EPA, the locations of the air quality are determined by boroughs instead of streets like in the Traffic Volume dataset, so it made it much harder to have a more accurate location in streets. This is another big limitation to have because having the location be a broad and generalized area like a borough makes it harder to have predictions of air quality in certain locations in a borough.

Evaluation

- Evaluate your model. Do you feel confident about its performance?

  For our model that we have with Traffic Volume vs Air Quality, there are problems with the model itself in terms of how it performed. We made a Linear Regression model and with our line, it went in a horizontal direction, which indicated a problem with the data. The performance of this model turned out to be poor because of how the datasets were made and we later found out that there really wasn't any correlation between the datasets,

hence our flat line in the model below. The plotting of the data was also heavily skewed on the left side of the graph.



Dep. Variable:	aqi	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.6978
Date:	Fri, 16 Dec 2022	Prob (F-statistic):	0.404
Time:	12:34:58	Log-Likelihood:	-48330.
No. Observations:	12058	AIC:	9.666e+04
Df Residuals:	12056	BIC:	9.668e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	33.9461	0.206	164.594	0.000	33.542	34.350
vol	-0.0014	0.002	-0.835	0.404	-0.005	0.002

Omnibus:	1586.884	Durbin-Watson:	2.025
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3106.223
Skew:	0.832	Prob(JB):	0.00
Kurtosis:	4.847	Cond. No.	209.

```
X_test = X_test.reshape(-1, 1)
y_pred = linreg.predict(X_test)

MSE = mean_squared_error(y_test, y_pred)
print(MSE)
RMSE = np.sqrt(MSE)
print(RMSE)
MAE = mean_absolute_error(y_test, y_pred)
print(MAE)
```

```
186.70587805576008
9.669938099671372
13.664035935833896
```

Looking at our performance metrics as well, from the R-Squared to the Root Mean Squared Error, we can see that for our R-Squared, our performance turned out to be 0, which means that the model isn't performing well at all. The higher the R-Squared implies a better model. For our Root Mean Squared Error, we achieved a score of 9.66, which also implies a poorer model as a lower score means the model is better for this metric.

Future Work

For the future, more datasets could be incorporated that provide a lot more information in terms of quantity over the years as we had one dataset that we removed due to insufficient amount of data covered over a longer period of time. Different types of models could further be explored and used as well to obtain higher accuracy scores and better prediction. We also broadened the scope. We also used the 5 boroughs as locations, which is a very broad scope, so for the future, it would be better to narrow it down into neighborhoods within each borough to make the data more precise to a specific area instead of a borough.

Due to the limitations in our dataset,we found from our Linear Regression model that there is little to no correlation between air quality and traffic volume because the amount of information we had after merging and dropping any values did not match. The data we had was also very generalized by boroughs as other data did not have a more specific geo location such as borough name. This made our data look like there was no correlation because there can be parts of boroughs that have less or more traffic and a different air quality than other neighborhoods. To combat this we can try working on datasets that are better at merging together with more relevant and specific datasets. It would be better if the data were collected as one dataset so we would not need to lose any information when combining the data leading to very little correlation.

Some interesting questions we uncovered while working on the model is whether traffic volume is actually a broad approach to finding the connection to air quality and if it's something more specific that leads to bad air quality. Overall, a lot more work could be done to improve upon this, but it was a great learning experience to see how important it is to obtain data relevant to our question and which parts of the datasets that need to be cleaned as well to obtain good accuracies and models.