Colorado Pollution Levels

A study of air pollution levels in Colorado over time

Statistical Question & Hypothesis

- Statistical Question: As a nation, we continue to become more aware of environmental issues like pollution. The government imposes environmental standards to help mitigate pollution and there has been a huge movement in both the public and private sectors, towards clean energy sources like electric, solar, and wind power. But are we making an impact?
- Hypothesis: My hypothesis is that pollution levels will decrease over time?

Air Quality Index Measurement

- AQI Measurement: U.S. Air Quality Index (AQI) standard measures air pollution rates based on the individual air quality indexes of six factors: ground-level Ozone (0_3), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2), and particulate matter ($PM_{2.5}$ & PM_{10}).

AQI Basics for Ozone and Particle Pollution

Daily AQI Color	Levels of Concern	Values of Index	Description of Air Quality
Green	Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Yellow	Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Orange	Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Red	Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Purple	Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.
Maroon	Hazardous	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

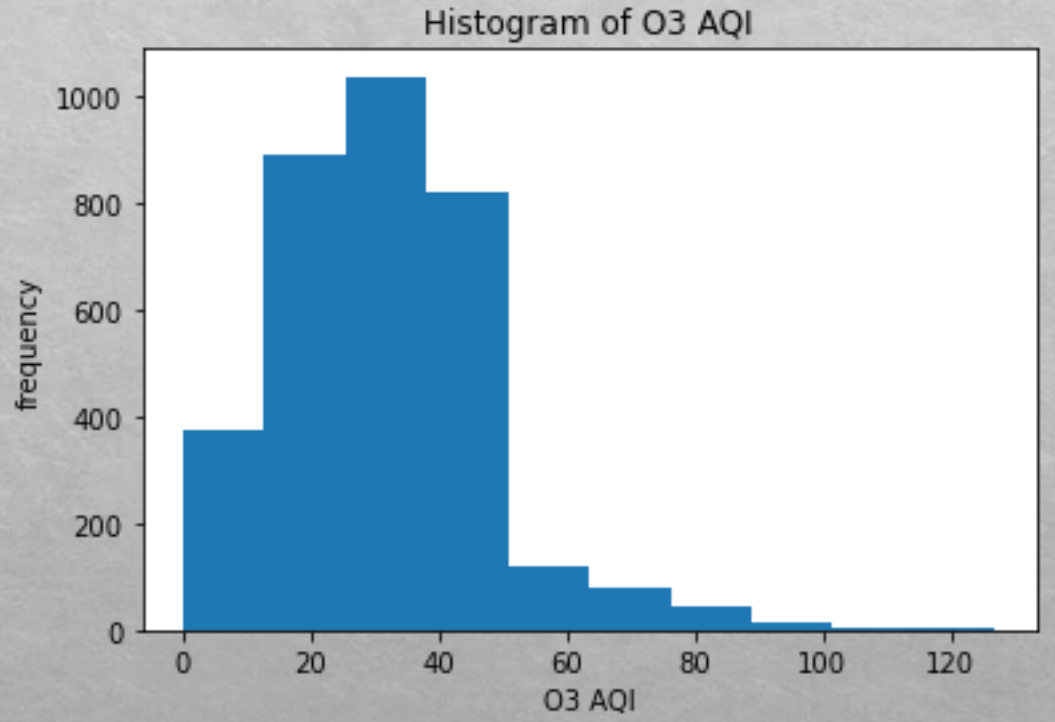
Figure 1: U.S. AQI Table (AirNow.gov, 2020).

Data Set Contents

- Daily Individual AQI measurements of:
  - Ozone (O3)
  - Carbon Monoxide (CO)
  - Sulfur Dioxide (SO2)
  - Nitrogen Dioxide (NO2)
  - Highest Daily IAQI
- Timeline: 1/1/2003 – 12/31/2010
- Retrieved from:

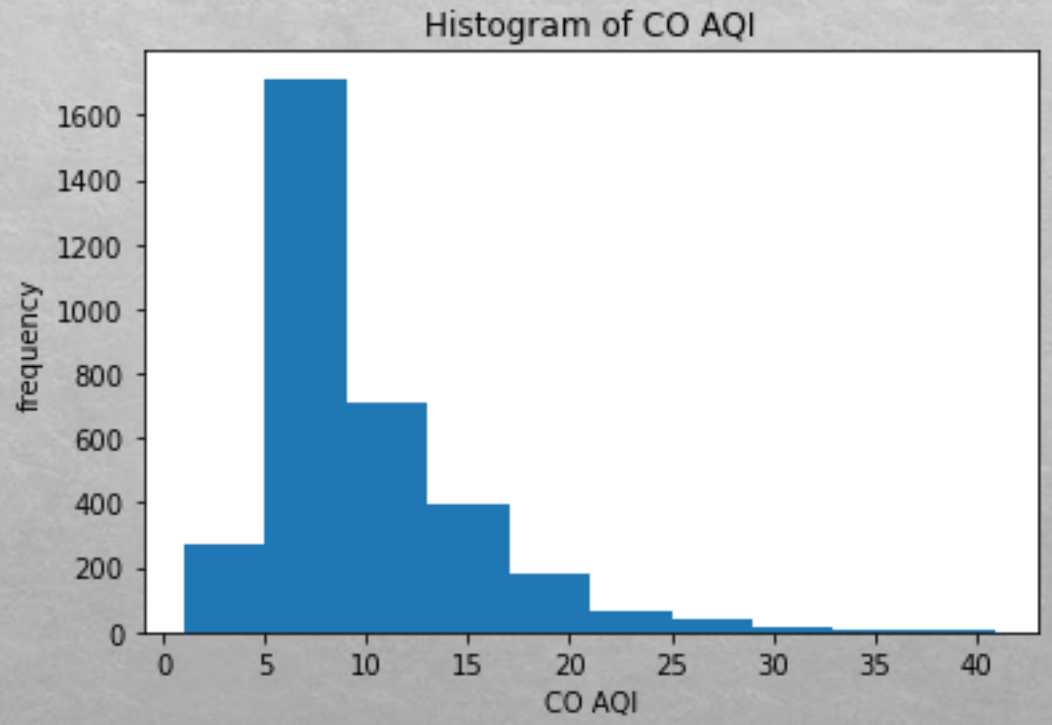
Histogram of Ozone IAQI

- Mean: 31.86
- Mode: 25
- Range: 0-127
- IQR: 20-42
- Variance: 268.76
- Skew: moderate positive skew 0.9
- Kurtosis: platykurtic 2.26



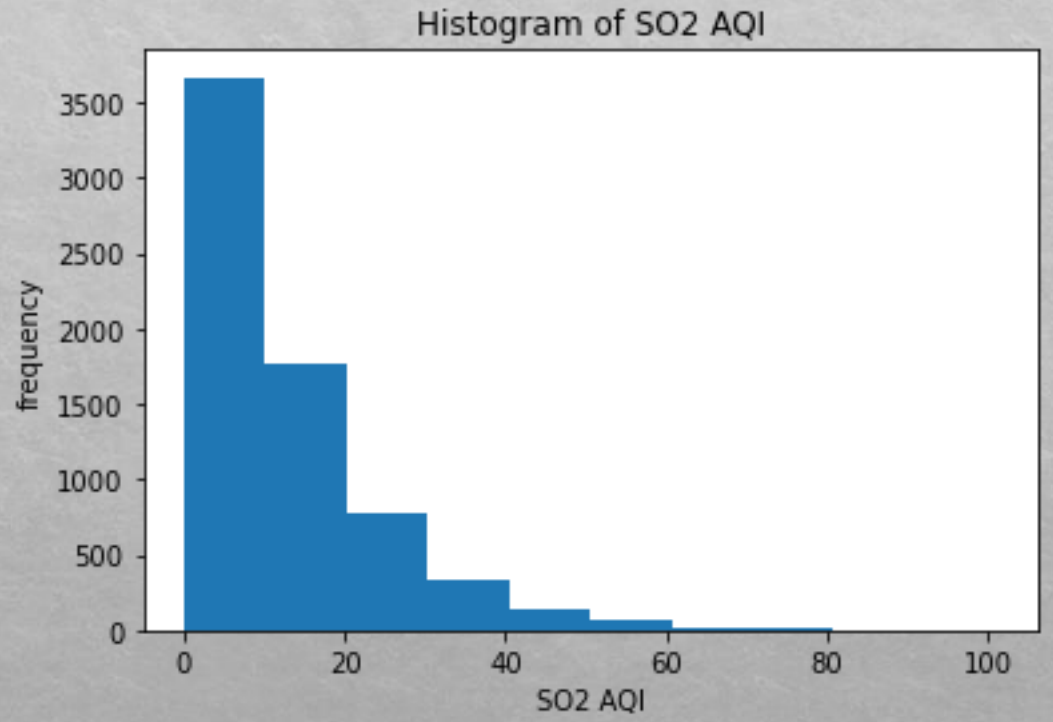
Histogram of Carbon Monoxide IAQI

- Mean: 9.16
- Mode: 6
- Range: 1-41
- IQR: 6-11
- Variance: 26.76
- Skew: high positive skew, 1.77
- Kurtosis: leptokurtic, 4.36



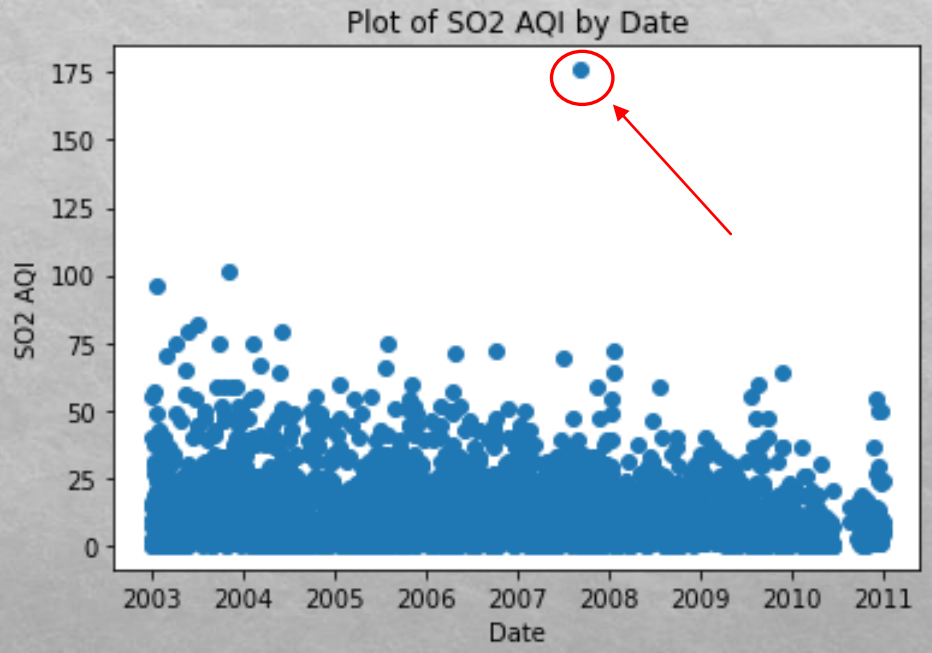
Histogram of Sulfur Dioxide IAQI

- Mean: 12.99
- Mode: 0
- Range: 0-101
- IQR: 4-19
- Variance: 148.73
- Skew: high positive skew, 1.78
- Kurtosis: leptokurtic, 4.54



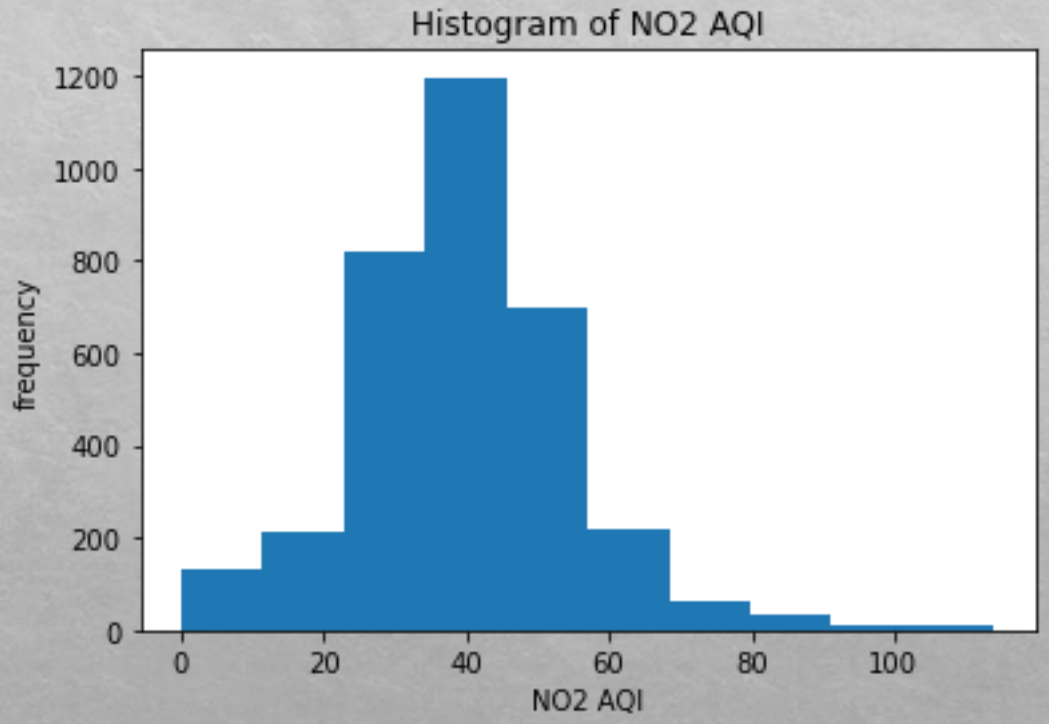
SO2 Outlier

- EDA revealed an extreme outlier in the SO2 variable.
  - Identified by scatterplot and very high kurtosis.
  - The SO2 value was 176, with the next closest value being 101.
- It is unknown what caused the outlier. SO2 AQI is produced by burning of fossil fuels and can be affected by forest fires, but this was only 1 day.
- Due to the extreme nature of this value and the possibility that it could degrade the value of the data, I have chosen to remove it.



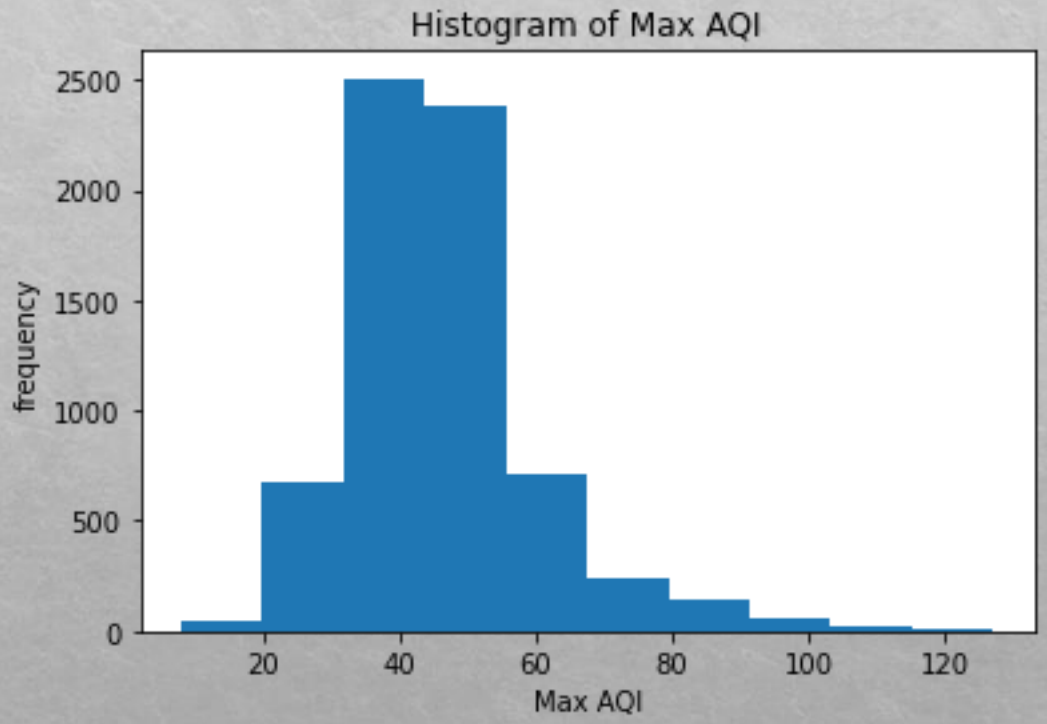
Histogram of Nitrogen Dioxide IAQI

- Mean: 39.63
- Mode: 42
- Range: 0-114
- IQR: 31-48
- Variance: 224.05
- Skew: Symmetrical, 0.46
- Kurtosis: platykurtic, 1.8



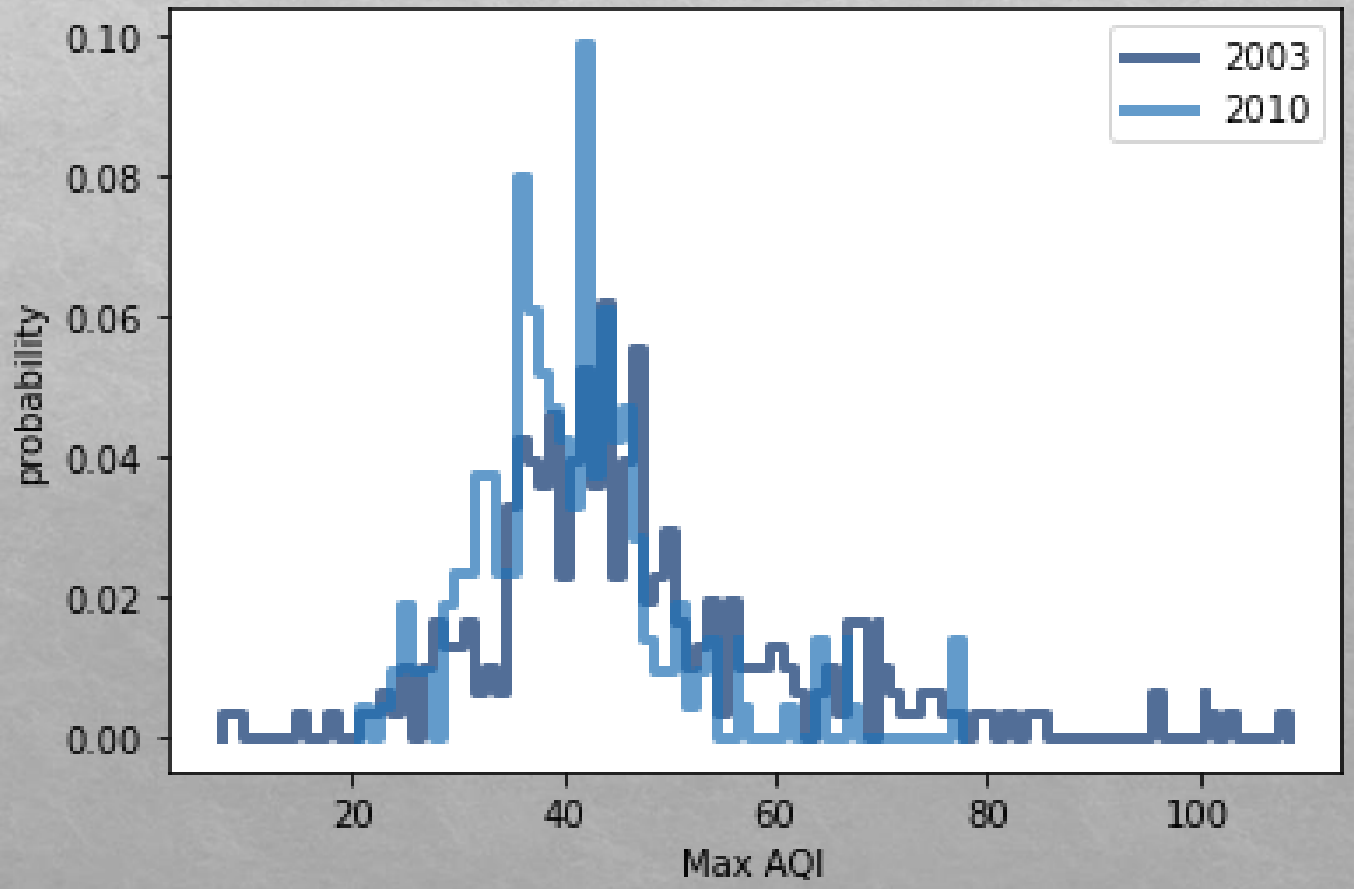
Histogram of Highest Daily IAQI

- Mean: 46.06
- Mode: 42
- Range: 8-127
- IQR: 38-51
- Variance: 191.38
- Skew: high positive skew, 1.3
- Kurtosis: leptokurtic, 3.33



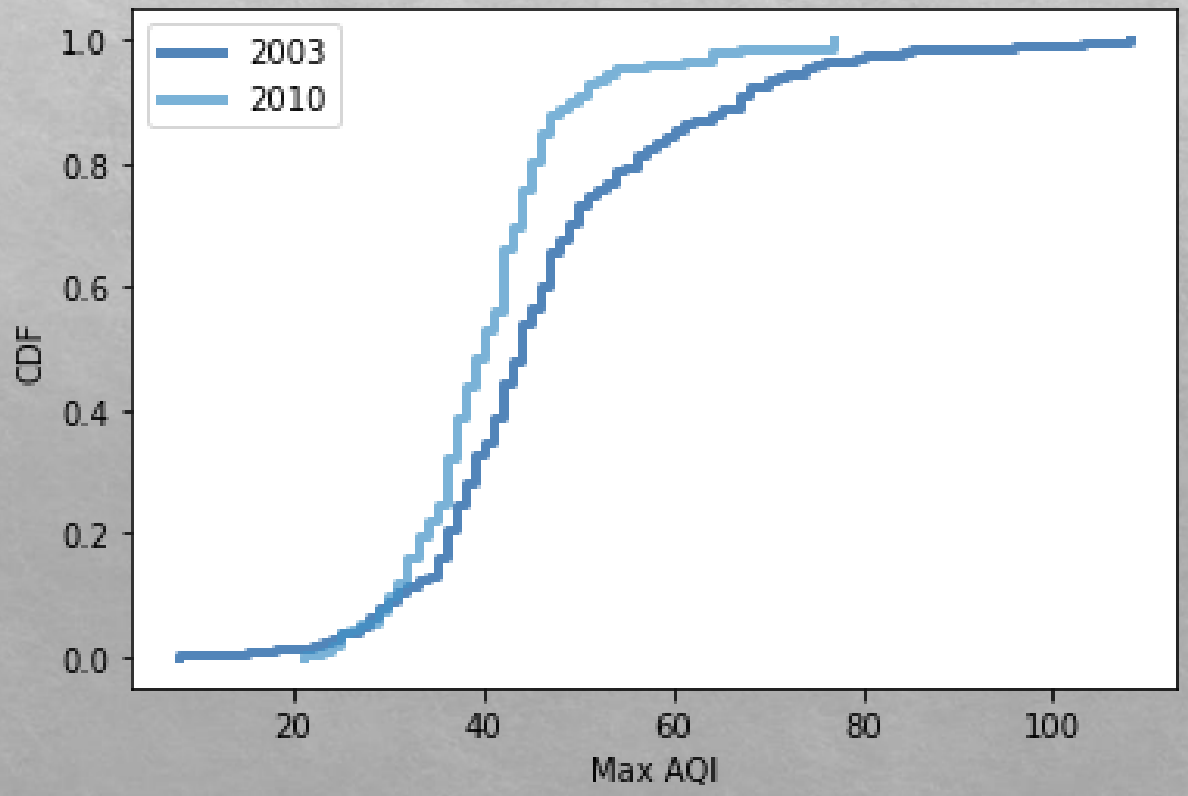
PMF

- Max AQI data from 2003 and 2010 to compare the PMFs.

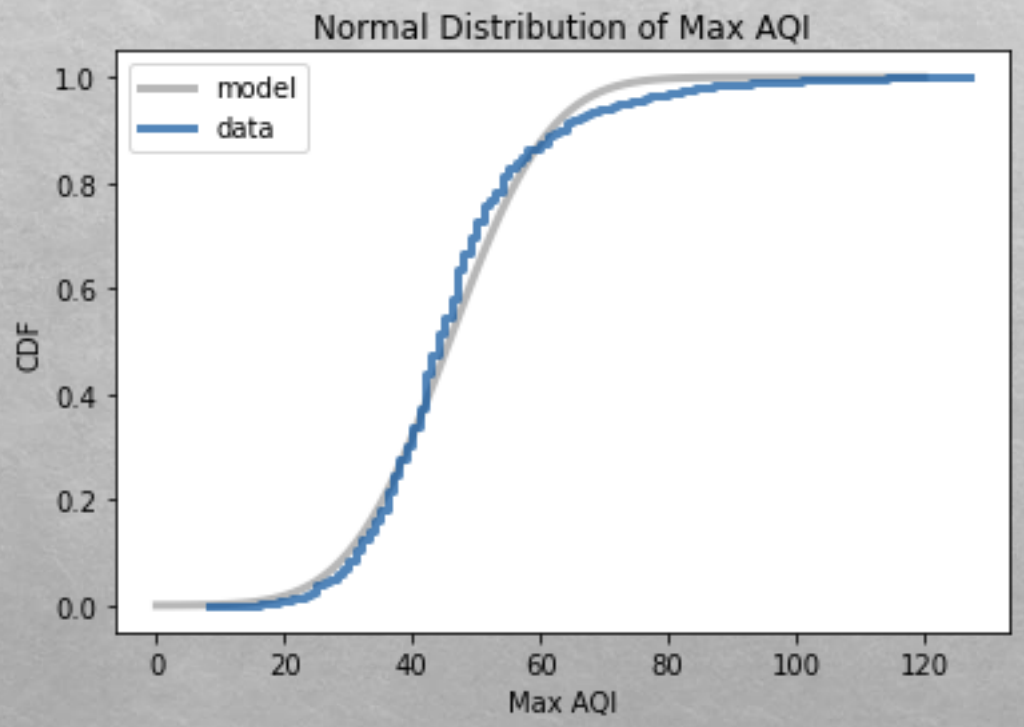


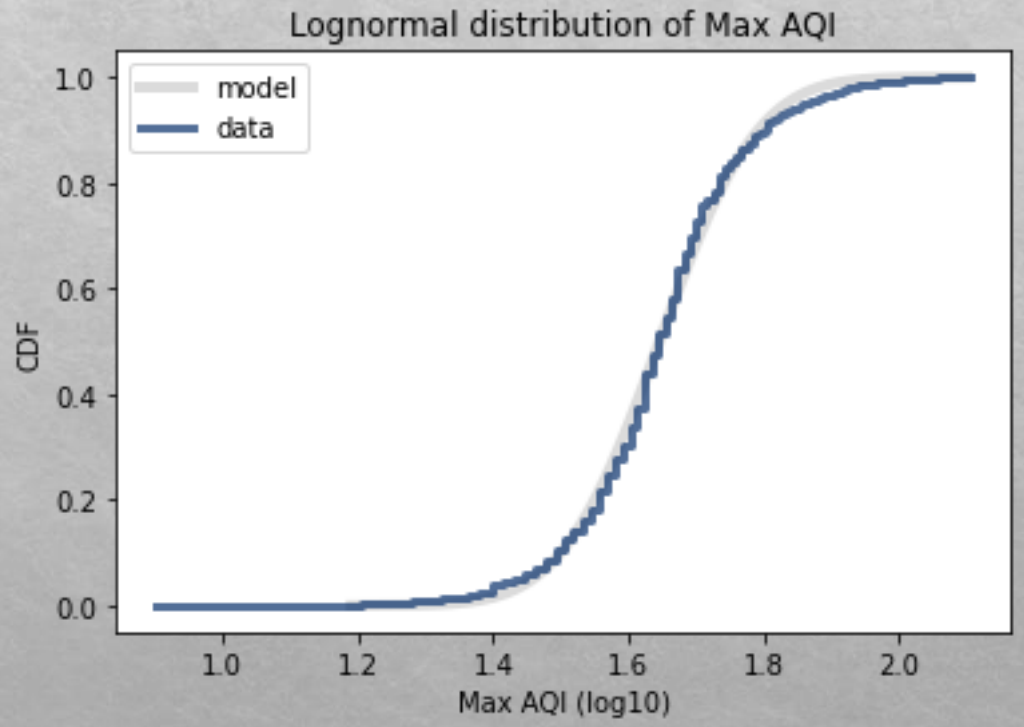
CDF

- PMF was difficult to gain insight from, so I used the same values to plot a CDF comparison.
- 2010 has lower Max AQI scores.
- The discrepancy gets considerably larger above the mean. 2003 had NO2 scores that were much higher than 2010.
- This indicates that my hypothesis could be correct: Pollution levels have decreased over time.



Normal vs Lognormal Distribution

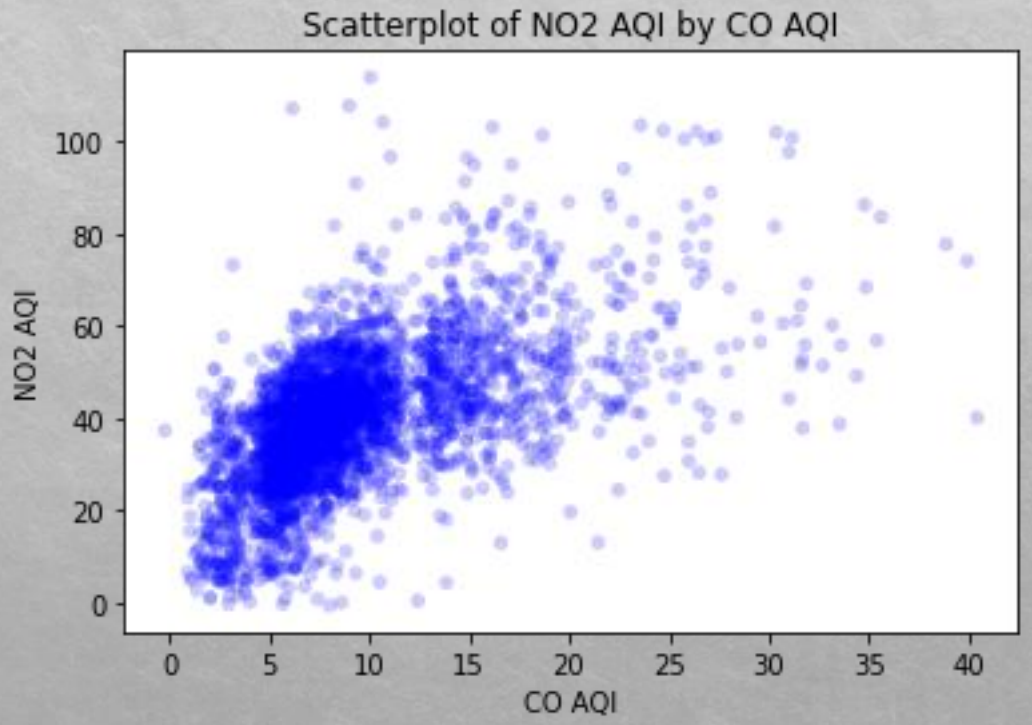




- The same distribution is shown on a linear scale and a log scale.
- The data fits the lognormal model better.

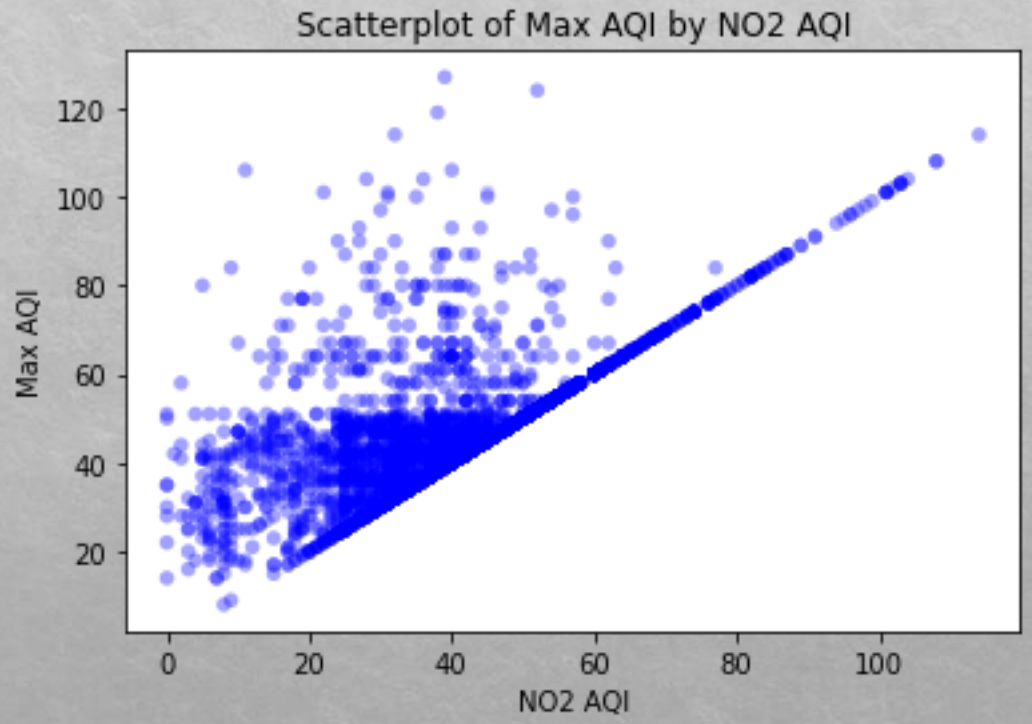
Scatterplot of NO2 AQI by CO AQI

- Positive relationship
- Covariance: 44.2
- Pearson's Correlation Score: 0.57
- Spearman's Rank Correlation: 0.61
  - Closeness suggests skewness in distribution explains most the difference between Pearson's and Spearman's correlation scores.



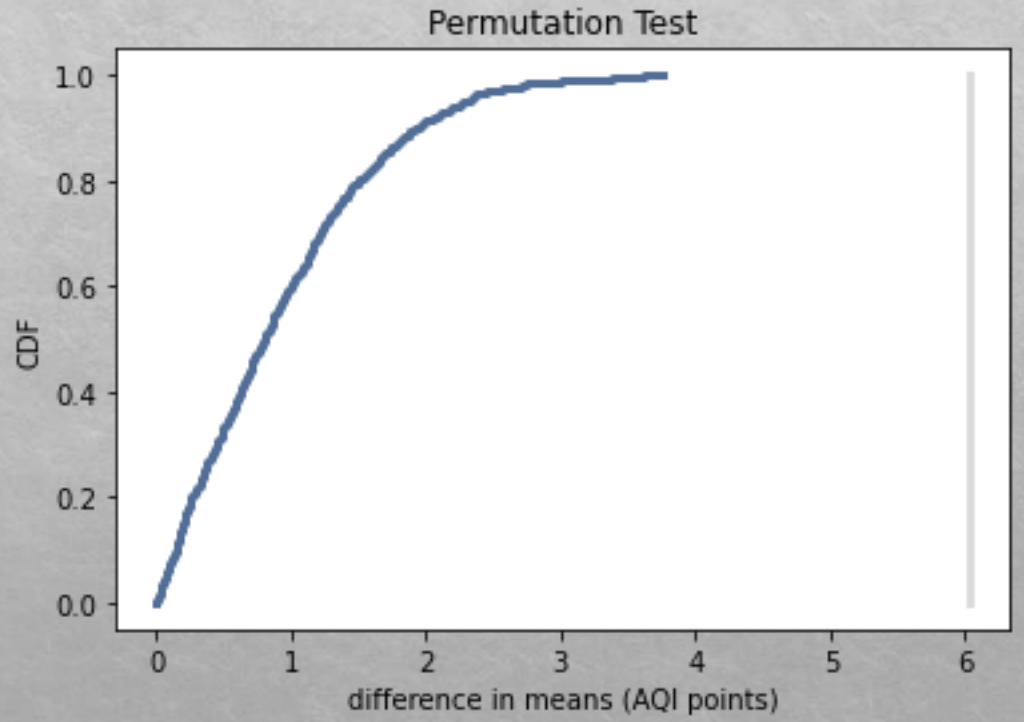
Scatterplot of Max AQI by NO2 AQI

- Positive relationship
- Covariance: 136.12
- Pearson's Correlation Score: 0.66
- Spearman's Rank Correlation: 0.69
  - Closeness suggests skewness in distribution explains most the difference between Pearson's and Spearman's correlation scores.
- Causation: Max AQI is the highest scoring daily AQI. NO2 is often the highest scoring, therefore NO2 can be considered to have a certain amount of causality on Max AQI.



Hypothesis Test

- Used permutation to test the difference in means between Max AQI in 2003 and Max AQI in 2010. The calculated difference in means was 6.03 points.
- Results:
  - After 1000 attempts, the simulation never yields an effect as big as the observed difference of 6.03.
  - $P < 0.001$, the difference in Max AQI scores is statistically significant and the null hypothesis is rejected.



Regression Analysis

- Linear regression results - Max AQI by Year:
  - Intercept: 47.166
  - Slope: -0.288
  - R2: 0.002
- Meaning:
  - Mean Max AQI decreases by 0.288 points per year
  - R2 of 0.002 indicates year is only accountable for 0.2% of observed variability in Max AQI

OLS Regression Results

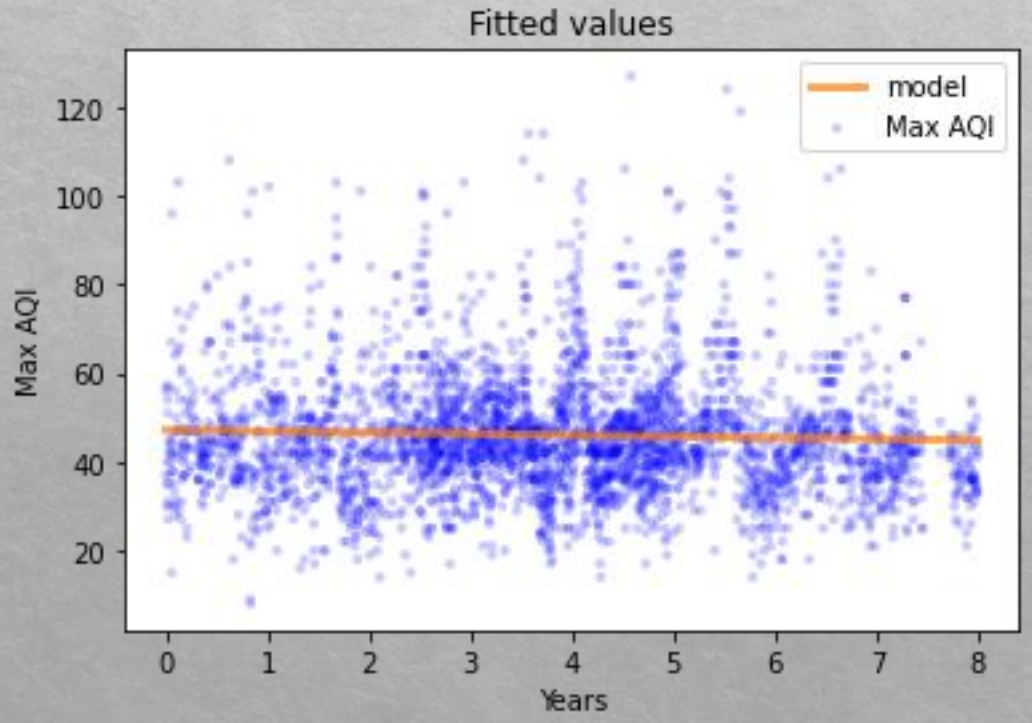
Dep. Variable:	maxAQI	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	0.001
Method:	Least Squares	**F-statistic:**	11.02
Date:	Sat, 21 Nov 2020	**Prob (F-statistic):**	0.000906
Time:	15:22:36	**Log-Likelihood:**	-27467.
No. Observations:	6790	**AIC:**	5.494e+04
Df Residuals:	6788	**BIC:**	5.495e+04
Df Model:	1		
Covariance Type:	nonrobust		

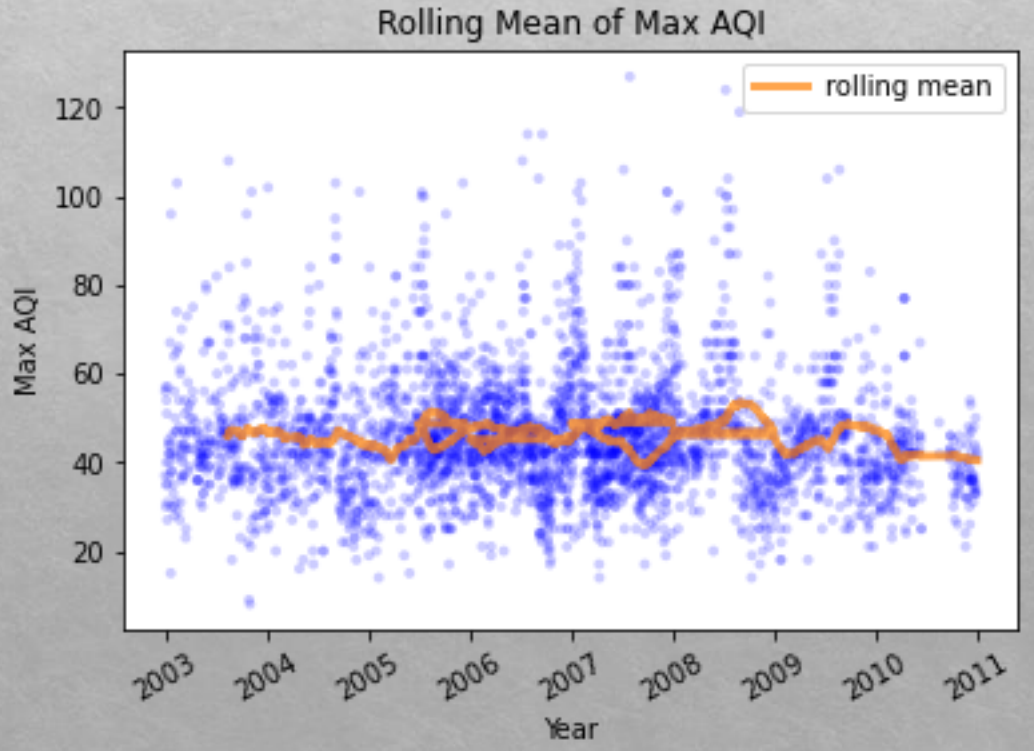
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	47.1657	0.373	126.614	0.000	46.435	47.896
Years	-0.2876	0.087	-3.320	0.001	-0.457	-0.118

Omnibus:	1706.838	**Durbin-Watson:**	0.583
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5104.216
Skew:	1.297	**Prob(JB):**	0.00
Kurtosis:	6.364	**Cond. No.**	9.96

Regression Analysis

- Due to time not commonly producing linear data, I also plotted the annual rolling mean





References

- AirNow.gov. (n.d.). AQI Basics. AirNow.gov. Retrieved from https://www.airnow.gov/aqi/aqi-basics/#:~:text=Think%20of%20the%20AQI%20as,300%20represents%20hazardous%20air%20quality.
- So, B. (2016, November 4). U.S. Pollution Data: Pollution in the U.S. Since 2000, Version 1. Retrieved September 27, 2020 from https://www.kaggle.com/sogun3/uspollution.