**Natural Determinants of Air Quality in Utah**

BMI 6106 Project Submission |Jessica Jung and Tricia Macauley

GitHub Repository:<https://github.com/tricia-ying/BMI6106-healthcarecost.git>

**Introduction**

According to theSalt Lake Tribune, Utah ranks among the worst places globally for air quality due to intersecting geographic, transportation and industrial factors (*Plothow, 2022*). Studies have shown that exposure to air pollution causes millions of premature deaths and results in the loss of years of life from otherwise healthy individuals (*​​Copenhagen and Geneva, 2021*), to the point where air quality has been classified as a health indicator by the World Health Organization. Population health is particularly impacted by the concentration of fine particulate matter (PM 2.5). While there is widespread news about the poor air quality in Utah, most articles and journals focus on measurements from the Salt Lake City-Provo-Orem area, which is ranked as one of worst cities for ozone and pollution by the American Lung Association (ALA).The organization, Your Utah, Your Future, states that, “For most of the year, Utah’s air is clean and meets federal health standards. But for some periods in the summer and particularly in the winter, regional weather patterns and topography, combined with emissions from various sources, cause poor air conditions along the Wasatch Front, in Cache Valley, and even in the Uintah Basin.”

We are interested in observing some underlying natural factors which have not been widely discussed, namely geographic location and seasonal climate, that contribute to air quality status across the state of Utah. Due to a land size of 89,899 mi², Utah houses a variety of geographical features, which can act as contributing factors for better or worse air quality. For example, Salt Lake County in northern Utah is prone to stagnant polluted air being left behind by inversions due to being surrounded by mountainous topography (*The Universe, 2017*); whereas Washington County in southern Utah is primarily characterized by the intersection of the Colorado plateau and Uintah Basin and is considered to have satisfactory air quality. Regional climate and seasonal patterns also impact air pollution levels, with indications that air quality is sensitive to temperature and can fluctuate even throughout the day with changes in temperature.

For this project, we will investigate the variables of geographic location and seasonal climate by conducting both a geospatial and longitudinal analysis on PM2.5 concentration and air quality index (AQI) values gathered throughout 2021 in the state of Utah. We will compare the annual average PM2.5 concentration and AQI values by county, as well as daily mean values over the entire calendar year to determine whether location or time has a greater effect on Utah’s air quality. With the results of our analysis, we will be able to explain the effect of these variables on air quality.

**Hypothesis**

*H₀:* Regional climate has a more significant impact on Utah air quality measurements than geographic location.

*Hₐ:* Regional climate does not have a greater impact than geographic location on Utah air

quality measurements.

*Assumptions*

In determining our hypothesis, we are accepting that the following is true based on research done prior to analysis:

1. Prior research on poor air quality in the SLC-Provo-Orem region is valid.
2. Utah has a documented seasonal weather pattern.
3. Man-made factors (e.g. emissions) are not the sole catalyst of poor air quality.
4. Utah has unique topography that can lead to natural occurrences, such as inversions, which worsen air quality.

**Data Methods**

*Source*

The primary data used within this analysis was obtained from the United States’ Environmental Protection Agency’s (EPA) Outdoor Air Quality database by querying for measurements in all Utah counties throughout calendar year 2021. Secondary data from Google Maps Platform was obtained for the use of geocoding using API to accurately plot air quality measurements over Utah topography for our geospatial analysis.

Utah AQI Data:<https://www.epa.gov/outdoor-air-quality-data/download-daily-data>

Google Maps Platform:<https://developers.google.com/maps>

*Description*

After reading the air quality dataset into R, we created a data frame with all columns from the *csv* file as listed below. Those shown in red are our objects of focus for this analysis.

Date, Source, Site.ID, POC, Daily.Mean.PM2.5.Concentration, UNITS, DAILY\_AQI\_VALUE, Site.Name, DAILY\_OBS\_COUNT, PERCENT\_COMPLETE, AQS\_PARAMETER\_CODE, AQS\_PARAMETER\_DESC, CBSA\_CODE, CBSA\_NAME, STATE\_CODE, STATE, COUNTY\_CODE, COUNTY, SITE\_LATITUDE, SITE\_LONGITUDE

Below are some descriptive statistics on our entire dataframe using the *describe* function in R.

Focusing on our columns of interest:

**Date** shows a minimum value of 1 and a maximum of 365, indicating that we have captured values for all days in the calendar year of 2021. These dates will be used as a primary independent variable for assessing air quality based on temporality and seasonality. This column is in the format of *mm/dd/YYYY*, and is continuous. As part of our processing, we elected to have air quality measurements consolidated into daily means for ease of manipulation, so the date values then became discrete measures that follow the previous value in unison.

**Daily.Mean.PM2.5.Concentration** is a continuous measurement given in units of micrograms per cubic meter (µg/m³) to assess concentration of fine particulates. This column contains values that will be our dependent variable in assessing air quality. The dataset shows that the PM2.5 concentration column has an overall annual mean of 8.26µg/m³, a standard deviation of 7.92, a minimum of -1.5000e+00µg/m³, and a maximum of 169.00µg/m³. There are multiple measurements taken at intervals throughout each day, so we elected to have the data consolidate these into daily mean measurements for AQI and PM2.5 for ease of processing and manipulation.

**DAILY\_AQI\_VALUE** is a continuous measurement and has an overall annual mean of 31.13µg/m³, a standard deviation of 23.45µg/m³, a minimum of 0µg/m³, and a maximum of 169.00µg/m³. This column contains values that will be our dependent variable in assessing air quality. There are multiple measurements taken at intervals throughout each day, so we elected to have the data consolidate these into daily mean measurements for AQI and PM2.5 for ease of processing and manipulation.

**COUNTY** is a categorical column which shows that air quality data was sampled from eleven distinct counties in Utah. This will be used as a primary independent variable for assessing air quality based on location. Looking further into which unique values appear, we can see that this includes the following counties:

Cache, Carbon, Davis, Duchesne, Iron, Salt Lake, Tooele, Uintah, Utah, Washington, Weber

**SITE\_LATITUDE** is a continuous variable and was included for geocoding and mapping air quality values onto a plot of Utah to assess the surrounding topography and geographical features. From the descriptive statistics, latitude has a mean of 40.59, a standard deviation of 0.92, a minimum of 3.7180e+01, and a maximum of 41.84.

**SITE\_LONGITUDE** is also a continuous variable and has a mean of -111.77, a standard deviation of 0.74, a minimum of 1.1331e+02, and a maximum of -109.56.

**Analysis Methodology**

*Data Processing*

In preparation, we reformatted and shaped our data frame to make it suitable for conducting our analysis. We grouped all measurements by date and averaged the daily value over POC to create a mean value for AQI measurement and PM 2.5 concentration each day in 2021. We also converted the data type for DATE from *character* to *date* for ease of manipulation.

*Exploratory Analysis*

We will conduct an initial analysis to determine the scope of our data, and whether all requirements for further analysis are fulfilled. This will also reveal some descriptive statistics to provide more insight on the structure of our data frame, and whether any visible trends are present.

*Geospatial Analysis*

We intend to use the SITE\_LATITUDE and SITE\_LONGITUDE values to create an overlay plot of air quality measurements on a real-to-life topography map of Utah. This will allow us to evaluate if air quality is visibly varied based on location.

*Longitudinal Analysis*

We intend to use the DATE values to plot daily air quality measurements over the year of 2021. This will provide an introspective view into major trends in air quality occurring throughout the year. In addition to this, we also will group all measurements by month to identify which periods bear significance in regards to poor air quality.

*Repeated-Measures ANOVA*

Since we selected our primary independent variables as time (DATE) and location (COUNTY), we needed to evaluate how to best assess each through a comparable statistical test. This presented difficulty because COUNTY, as a categorical value, would be simple enough to apply a variety of statistical models; however, DATE in its pre-existing structure would be difficult to implement. Most longitudinal analysis models focus on either cause-and-effect on time-series data or require a longer reporting timeframe (e.g. several years) to assess trends and seasonality, however our data only consisted of 2021 measurements. Therefore, we found a way to separately implement a Repeated-Measures ANOVA on each predictor variable by coercing DATE values into nominal values grouped by month, while still treating COUNTY as a categorical variable. The Repeated-Measures ANOVA test would allow us to compare the means and/or variance in each subset of our data based on repeated observations within each variable. Our objective with selecting this model was to determine whether time or location had a greater effect on the variance in means to provide insight on its overall impact on air quality.

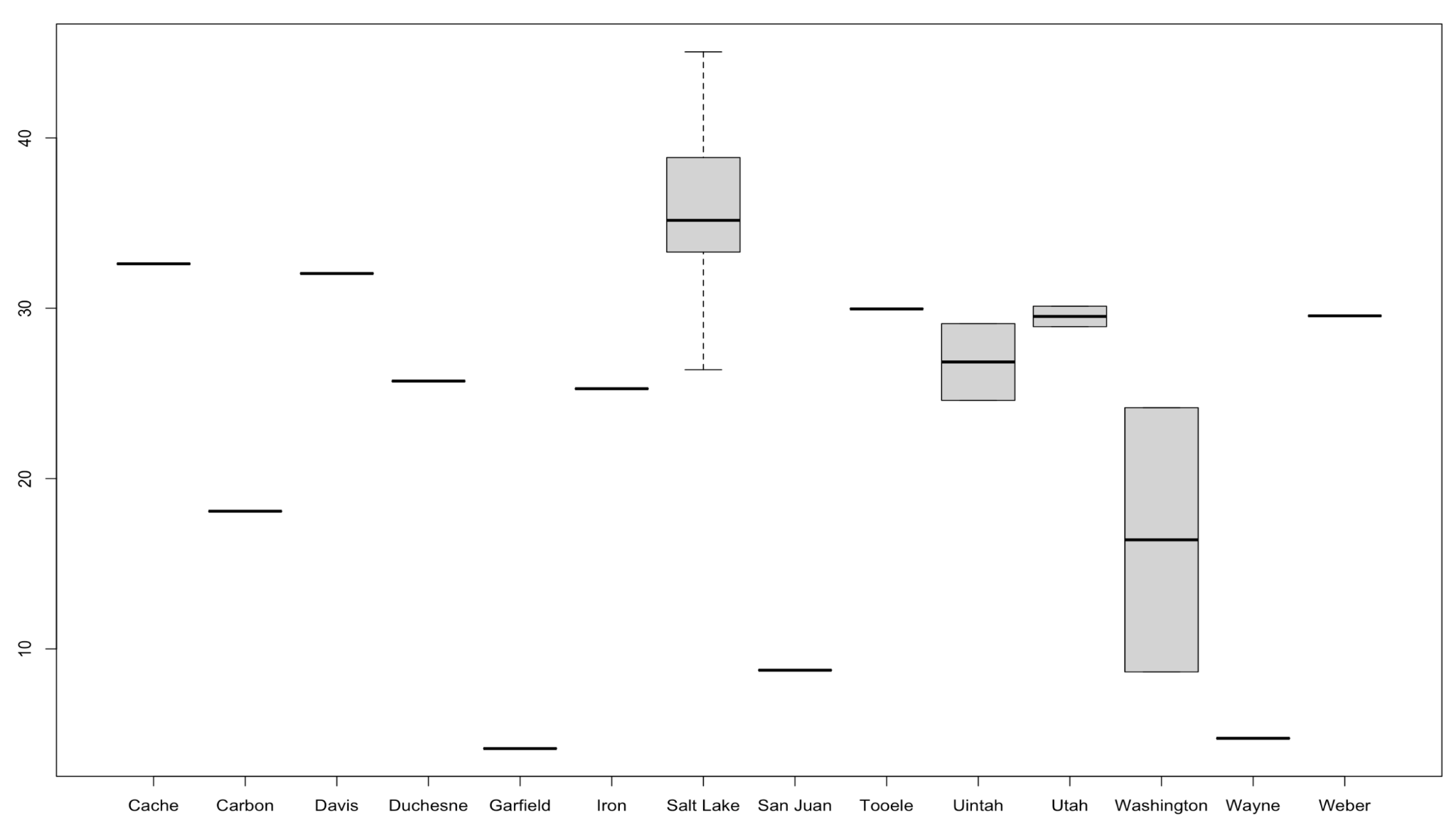
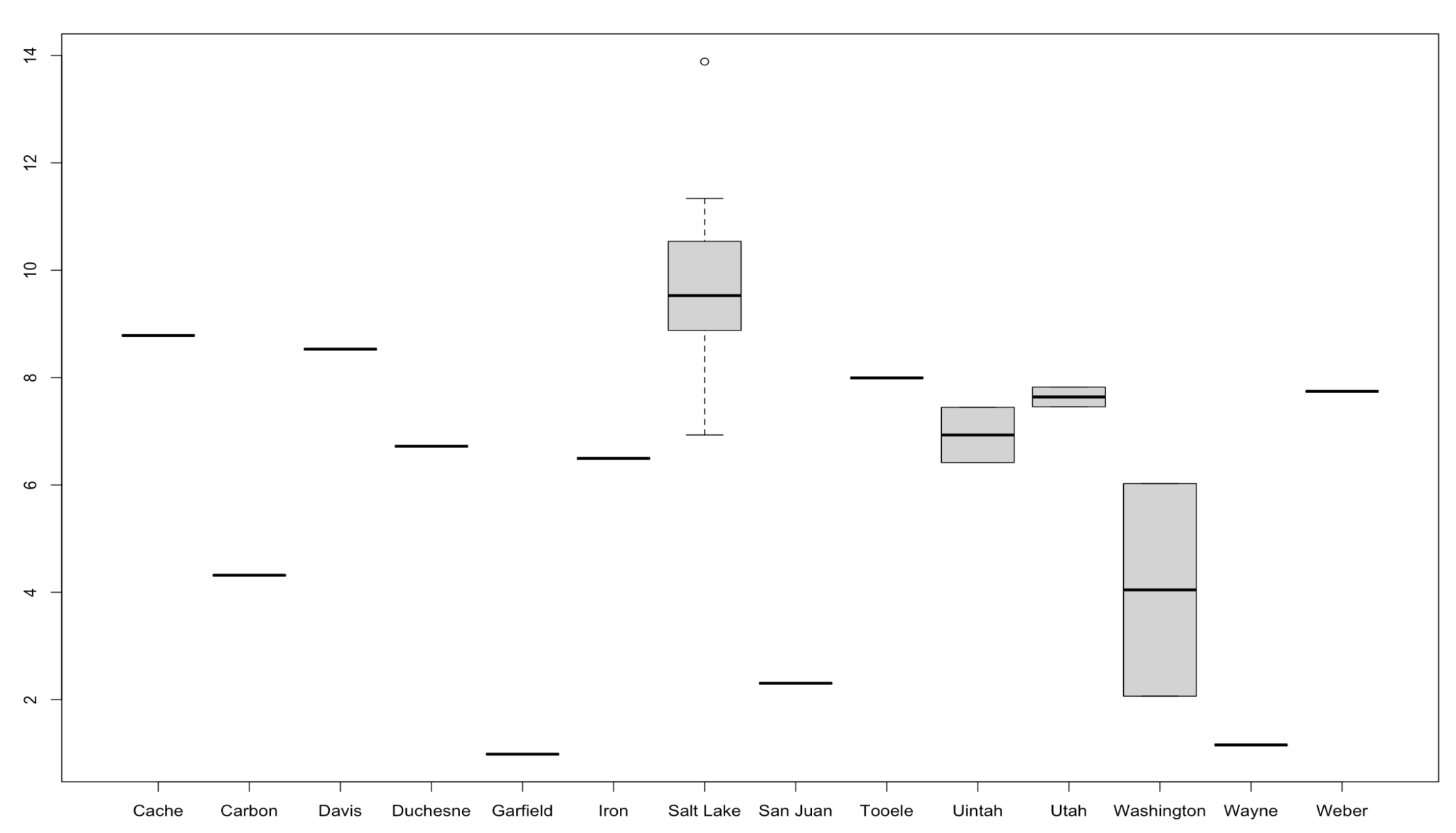
**Results**

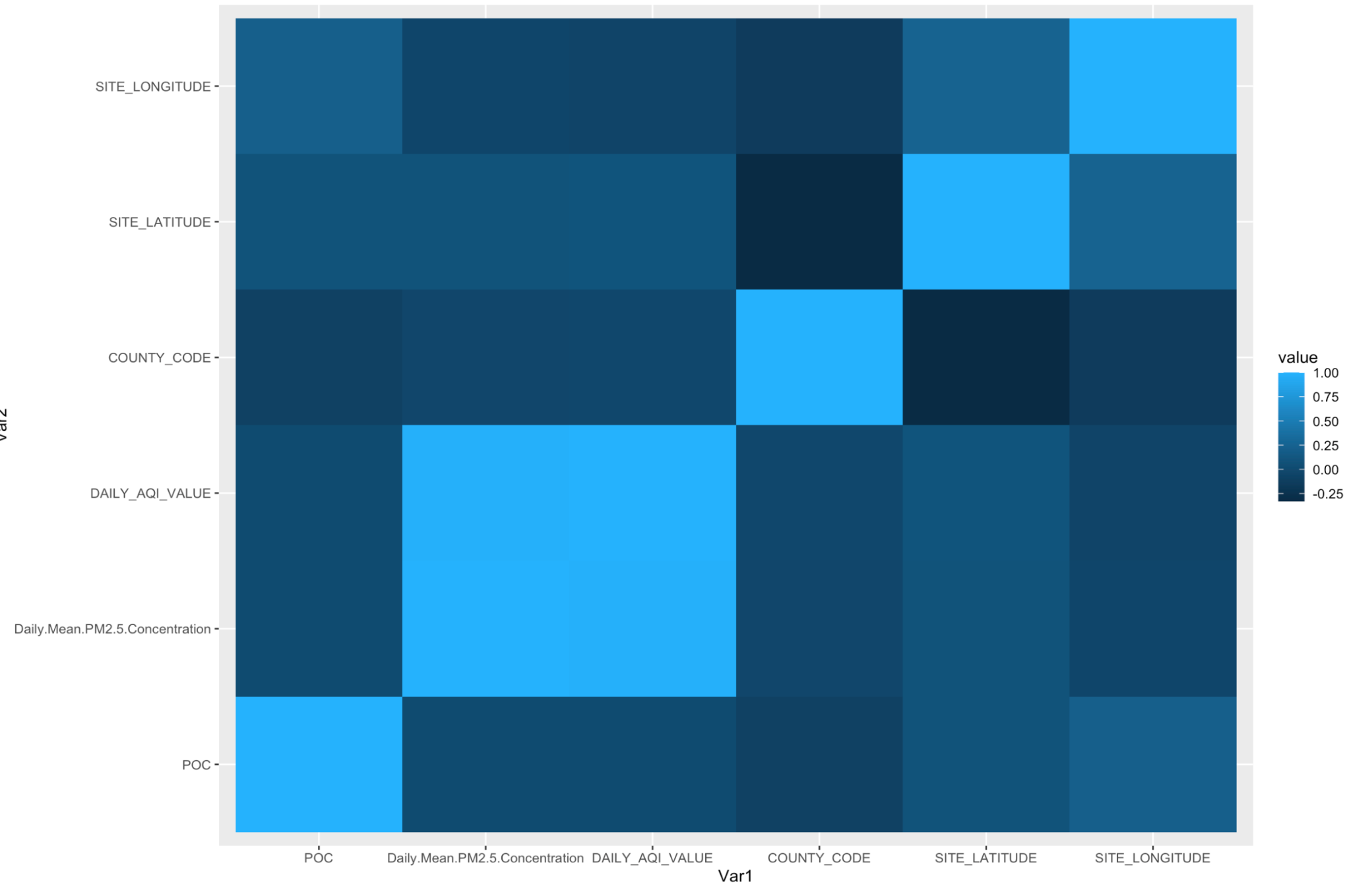
*Exploratory Analysis*

Our original data set had dimensions of: 20 columns by 11963 rows, with 100% completion. The overall mean for daily PM 2.5 concentration was 8.21, with a standard deviation of 7.91; the overall mean for daily AQI measurements was 30.94, with a standard deviation of 23.46. The standard deviation relative to the mean for both of these variables indicates that there is a large amount of variation in the data between values.

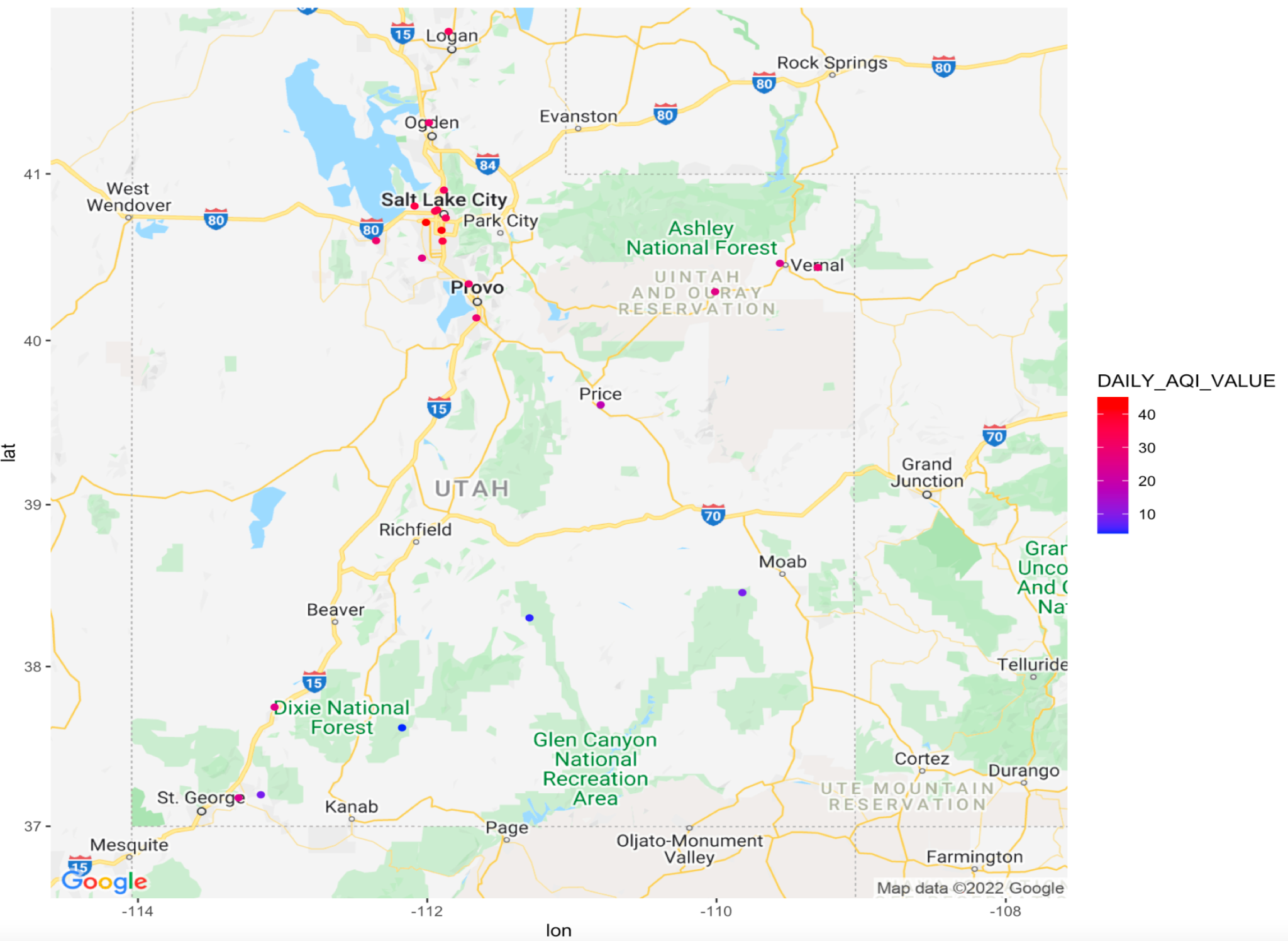
After grouping the measurements by county, we were able to create the boxplots below for an initial look into the proportion and ranges of air quality measurements. Most counties appear to have a very narrow range of values, with the exception of Salt Lake, Uintah, and Washington. While it is assumed that PM 2.5 concentration and AQI measurements should be similar, it is helpful to be aware that our dataset follows the same pattern.

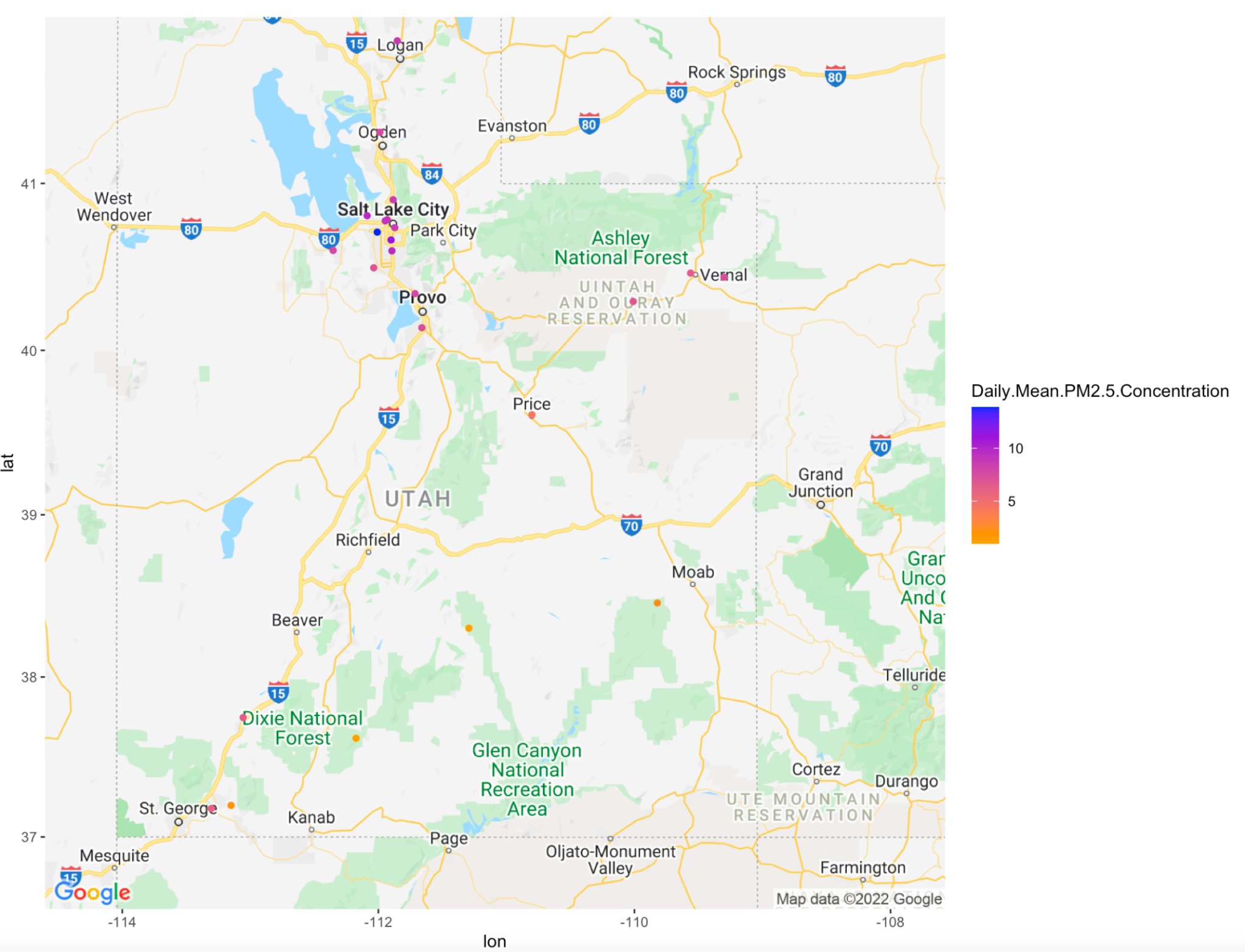
Daily AQI Values by County Daily PM 2.5 Concentration by County

Following this, we wanted to have a look into whether any of our variables were correlated. We plotted the below correlation matrix, where we can see that DAILY\_AQI\_VALUE and Daily.Mean.PM2.5.Concentration are highly correlated as expected. There also appears to be some weak correlation between SITE\_LATITUDE with DAILY\_AQI\_VALUE and Daily.Mean.PM2.5.Correlation. No other noticeable associations are visible. 

Based on the results of this initial analysis, we are able to confirm that all the required data for analysis is present, and that the data appears to be realistic without an abundance of outliers.

*Geospatial Analysis*



We superimposed the annual means of both the AQI measurements and PM 2.5 concentration over Google Platform’s topography map of Utah.

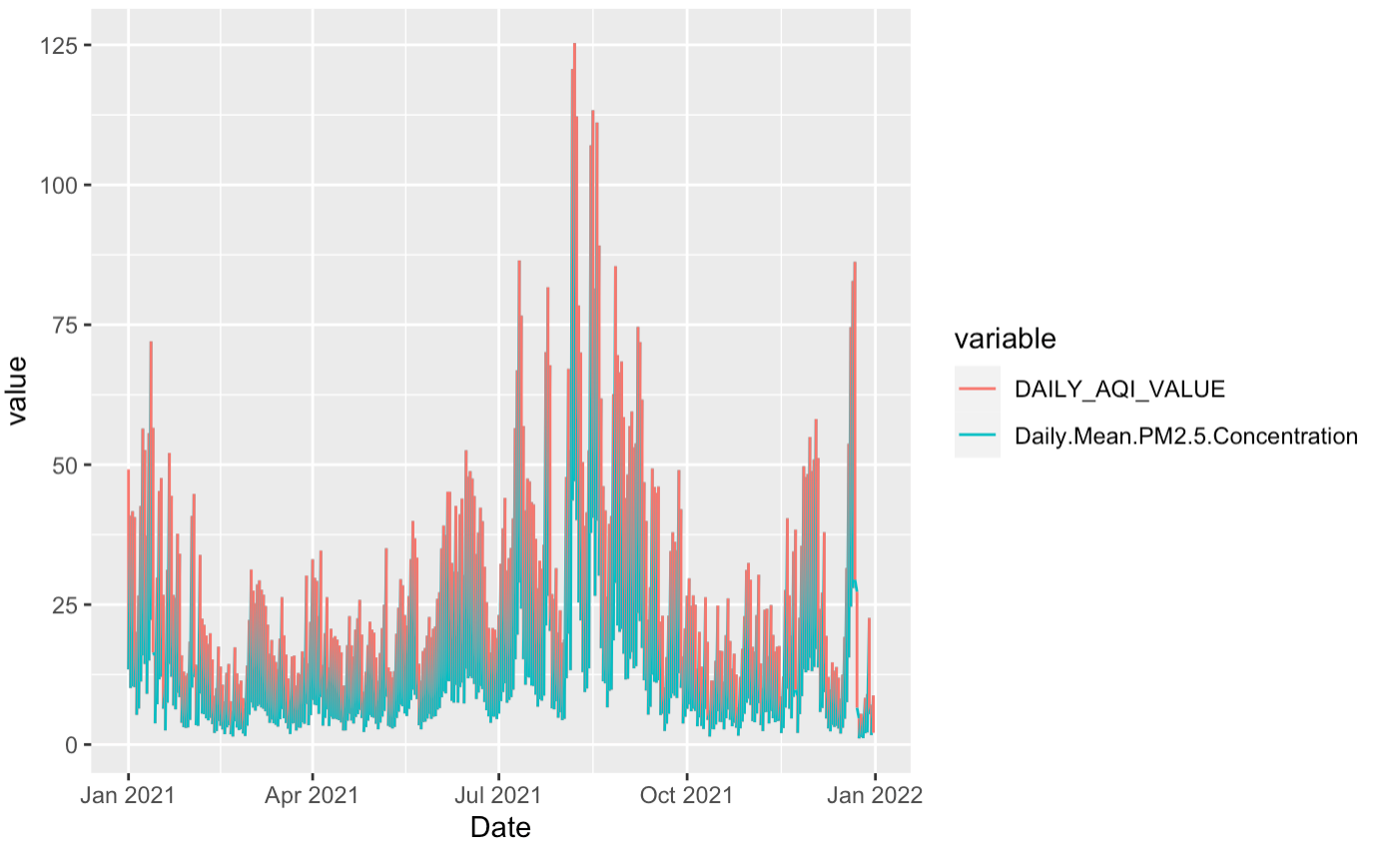
Both plots show the same patterns in air quality. As seen below, the annual means taken from the daily AQI values are higher at around 30-40µg/m³, in the Salt Lake-Provo-Orem area as we expected from our prior research. Southern Utah has significantly lower AQI values in the range of 10-20µg/m³.

This also appears to be true for our PM2.5 concentration values, with the Salt Lake City-Provo-Orem area having a range of >10µg/m³, whereas southern Utah counties fall in the <5µg/m³ range. The overall variance of mean AQI values by county is around 117.82, and the variance of mean PM 2.5 concentration by county is an estimated 9.87.

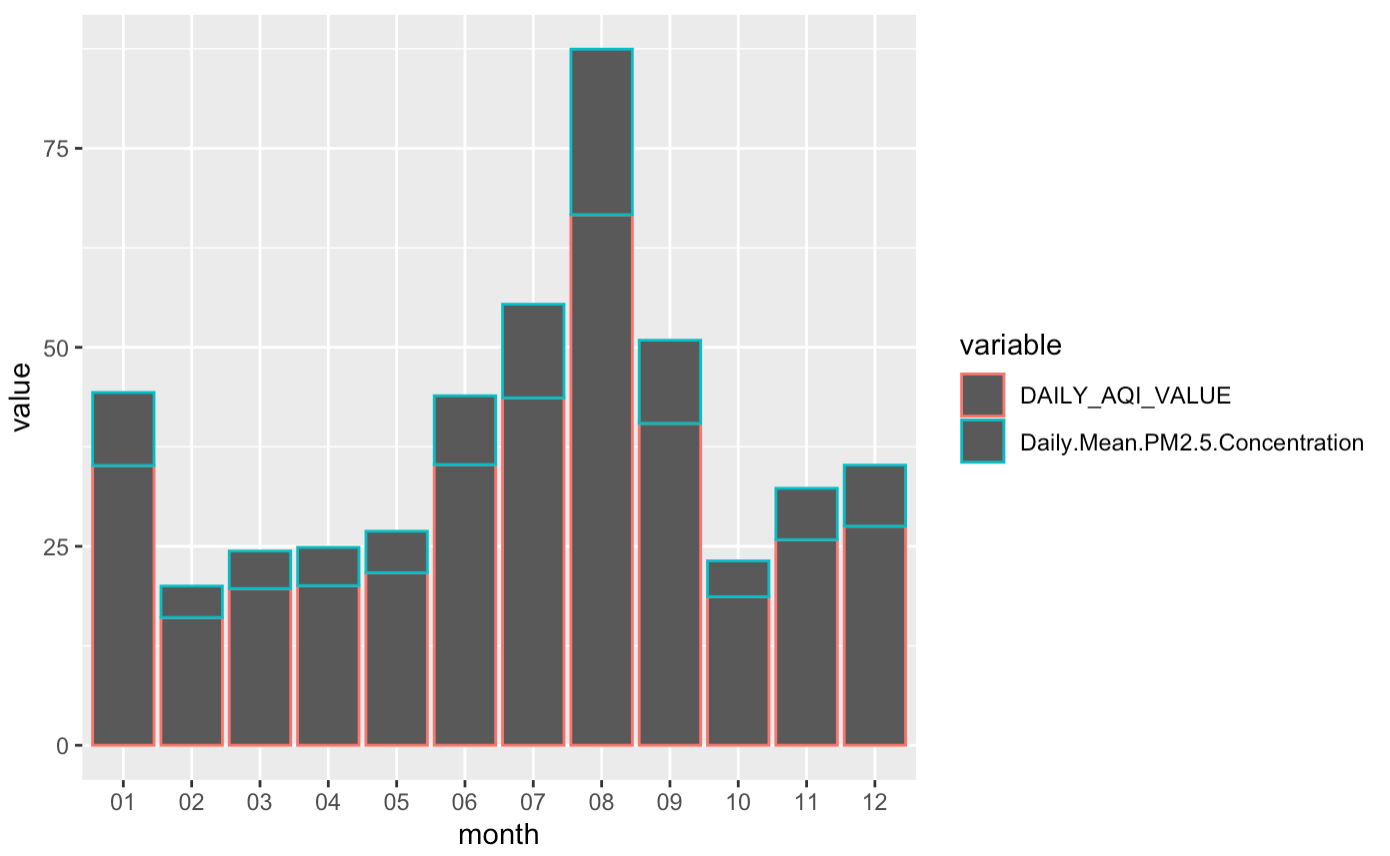
These visualizations lend evidence to geography having an impact on air quality, but we cannot determine the significance of association. In addition, the variance does not serve as a strong indicator on its own, but it can be compared with the longitudinal variance to provide initial insight into which subset may have greater variation between occurrences prior to implementing the Repeated-Measures ANOVA.

*Longitudinal Analysis*

From prior research, we are aware that it is common for the days with the worst air quality to fall in late summer due to the extreme heat, and also in midwinter due to inversions. From our time-series plot of daily air quality values below, we can see that the peak is right around August-September when temperatures tend to be the highest. This visualization also appears to be bimodal, with a second less amplified peak appearing in December-January.



To observe more rigid trends in the data, we have grouped the values as means by month. Here we can see the same pattern as above, with greater emphasis on August and January having the poorest air quality in 2021.



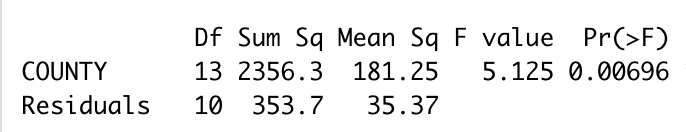
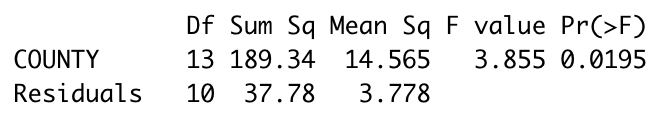
The variance in means by month for AQI measurements is approximately 209.46, and the variance for PM 2.5 is around 22.31. These measures are greater than those calculated based on a geospatial context, which indicates that there are more significant differences in air quality month-to-month than by county. While it is possible to identify the major trends for Utah’s air quality in 2021 from these plots and comparison of variances, we will need to apply ANOVA to observe further results.

*Repeated-Measures ANOVA*

Following this, we implemented the Repeated-Measures ANOVA model to solidify the significance of our findings. Evaluation of our model will use an 𝛼 = 0.05, which needs an F-value > 2.7 in accordance.

Beginning with geographic location, we calculated the following results for ANOVA:

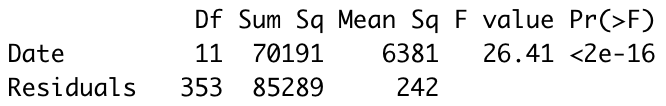
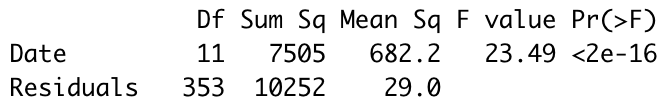
ANOVA: DAILY\_AQI\_VALUE ~ COUNTY ANOVA: Daily.Mean.PM2.5.Concentration ~ COUNTY

The p-value for AQI measurements is 0.00696, and for PM 2.5 concentration is 0.0195. These are both less than alpha, so we find that the relationship is statistically significant. In addition to this, the F value for both measurements is above 2.7 so the variation in means is substantiated. A significant relationship between geographic location and air quality was expected based on research prior to the analysis, but this does not yet provide evidence of one variable outweighing the other in significance.

Following this, we applied the ANOVA model to our monthly data for the following results:

ANOVA: DAILY\_AQI\_VALUE ~ monthly\_AQI2 ANOVA: Daily.Mean.PM2.5.Concentration ~ monthly\_PM2.52

The p-value for both AQI measurements and PM 2.5 concentration is <2e-16, which is far smaller than 0.05. The F value for both measures is large as well, which indicates that the model demonstrates a strong relationship between air quality and time of year (by month). The variation in means month-to-month appears to be substantially greater than by county.

**Conclusion**

Based on the findings in the geospatial and longitudinal analysis, as well as the Repeated-Measures ANOVA result, we are unable to reject the null hypothesis. There is strong evidence that regional climate has a more significant impact on Utah air quality measurements than geographic location. While both the geospatial and longitudinal plots show clear trends in air quality differences, comparing the variances and results of the Repeated-Measures ANOVA p-values shows that time of year has a greater impact on the variation of means in the data.

While our analysis has a clear outcome, we recognized that there are limitations in our dataset which made application of a statistical test difficult. For example, we only queried data for one calendar year (2021), assuming it would be sufficient. Most longitudinal studies assess stationarity or seasonality of data, which could not be established using our data set since climates require several years of data to establish a seasonal pattern. It is also possible that Utah’s land size is not large enough to constitute a difference in air quality, and geographic location could have a more significant impact on a larger scale. Our results as a whole could be solidified with a larger scope of analysis on both region and time.

In addition to this, overlap in impact may exist between location and climate; therefore, our predictor variables could also have a role as confounding variables. We attempted to minimize the interference of these variables with each other by taking annual measures for any geospatial data, and state-wide values for any longitudinal data. However, this realization also leads us to conclude that there are likely other confounding variables that act as catalysts for differences in air quality.

Overall, our analysis yields clear results that Utah’s air quality in 2021 was more significantly impacted by regional climate than by county. We believe our findings are substantive enough to apply to further analysis to solidify or dispute the results based on scope of data.

**References**

Copenhagen , & Geneva. (n.d.). *NEW WHO global air quality guidelines aim to save*

*millions of lives from Air Pollution*. World Health Organization. Retrieved February 15, 2022, from https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

HONOLULU, H. I. | A. 24. (n.d.). *Hawaii boasts some of the best air quality in the nation*

*according to the 'state of the Air' 2019 report*. American Lung Association. Retrieved February 15, 2022, from https://www.lung.org/media/press-releases/hawaii-boasts-some-of-the

News, F. (2021, October 18). *These are the US cities with the Worst Air Pollution*. New

York Post. Retrieved February 15, 2022, from https://nypost.com/2021/10/18/these-are-the-us-cities-with-the-worst-air-pollution/

Plothow, B. (2022, February 9). *It's a perfect time to improve classroom air quality in*

*Utah*. Retrieved February 15, 2022, from https://www.sltrib.com/opinion/commentary/2022/02/09/brad-plothow-its-perfect/

-, M. K. P., By, -, McKenna ParkMcKenna Park is an aspiring journalist studying at

Brigham Young University., Park, M. K., & McKenna Park is an aspiring journalist studying at Brigham Young University. (2018, February 12). Utah cities top charts for Worst Air Quality. The Daily Universe. Retrieved March 30, 2022, from https://universe.byu.edu/2017/02/28/poor-air-quality-affects-utahs-health1/

*Most polluted cities: State of the Air*. State of the Air | American Lung Association. (n.d.).

Retrieved March 30, 2022, from https://www.lung.org/research/sota/city-rankings/most-polluted-cities

*Your Utah, your future - background: Air Quality in Utah*. Your Utah, Your Future - Home.

(n.d.). Retrieved March 30, 2022, from https://yourutahyourfuture.org/topics/air-quality/item/44-background-air-quality-in-utah