Predicting NC Air Quality Index

Team 4
Angela Arce
Tammy Geis
Hanita Patel
Spencer Pope

August 25, 2022

Table of Contents - Segment 2 Deliverable

Predicting NC Air Quality Index (AQI)

The WHY

- Assist people with respiratory illnesses to determine if safe to engage in outside activities
- Information for people/families moving to NC to determine which region may best suit respiratory medical needs

The HOW

- Using various tools and Kaggle dataset predict AQI in regions across NC based on time of year
- App based tool for ease of use in Phase 2

The Data Questions

Does Air Quality vary by time of year?

What AQI is safe/unsafe for the respiratory system?

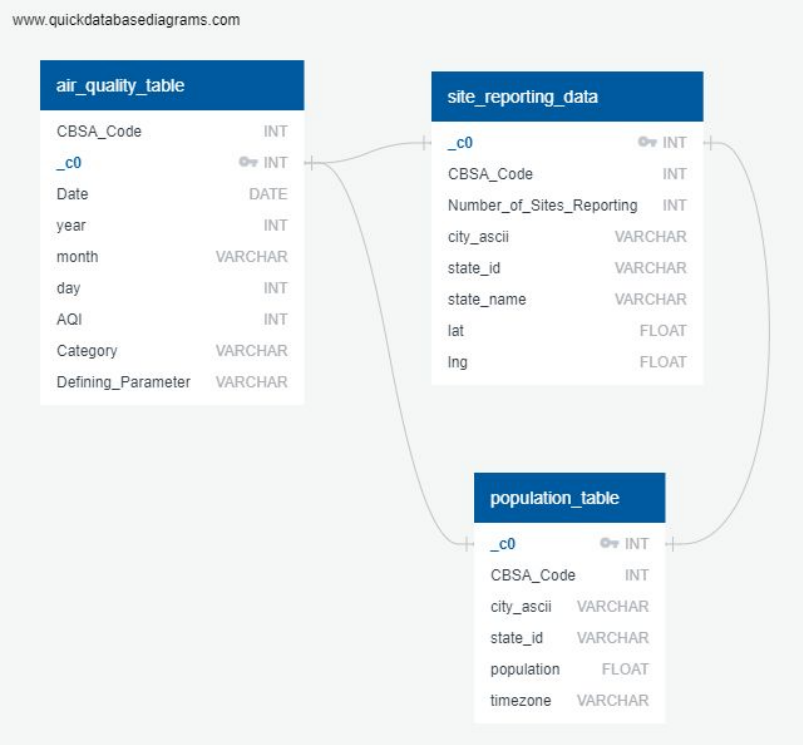
Does population density have an effect on AQI?

Does location have an effect on AQI?

What region of NC best suits someone with respiratory illness?

The Dataset

“US Air Quality 1980 - Present: Daily AQI Values from stations across the US” Source: Kaggle



Isolated data for NC from dataset to import to database as shown in database schema

Database ETL Details

AWS RDB and S3 bucket created to store original csv data file

Google Colab ETL File used Pyspark to create data frame

Data frame schema was updated to align with database ERD

Data frames were created for the tables: AirQuality, Site Reporting and Population

Data frames were written to the data base in Postgres Sql

Sample Code

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Team4-Project").config("spark.driver.extraClassPath","/conte
```

```
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://geisteam4-project.s3.amazonaws.com/ncaqi.csv"
spark.sparkContext.addFile(url)
user_data_df = spark.read.csv(SparkFiles.get("ncaqi.csv"), sep=",", header=True, inferSchema=True)
```

```
# Configure settings for RDS
mode = "append"
#jdbc_url="jdbc:postgresql://geisteam4.coe2ggfhl77s.us-east-1.rds.amazonaws.com"
jdbc_url="jdbc:postgresql://geisteam4.coe2ggfhl77s.us-east-1.rds.amazonaws.com:5432/postgres"
config = {"user":"postgres",
          "password": "xx",
          "driver":"org.postgresql.Driver"}
```

```
# Write airquality_df to table in RDS
airquality_df1.write.jdbc(url=jdbc_url, table='air_quality_table', mode=mode, properties=conf
```

```
# Write site_reporting_df to table in RDS
site_reporting_df1.write.jdbc(url=jdbc_url, table='site_reporting_table', mode=mode, properti
```

```
# Write site_reporting_df to table in RDS
population_df1.write.jdbc(url=jdbc_url, table='population_table', mode=mode, properties=confi
```

Machine Learning Details

Data Analysis Process

Does Air Quality vary by time of year?

What AQI is safe/unsafe for the respiratory system?

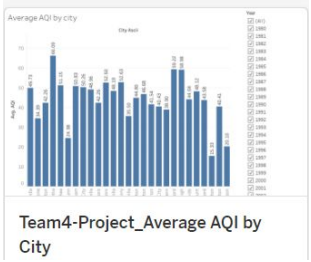
Does population density have an effect on AQI?

Does location have an effect on AQI?

What region of NC best suits someone with respiratory illness?

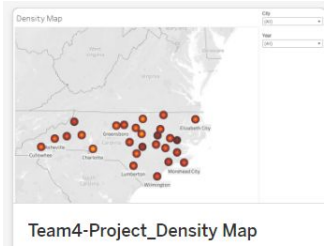
Data Visualization StoryBoard

Average AQI by City
Purpose to show AQI based on population of city

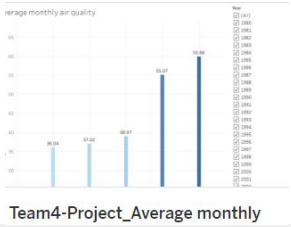


Annual Average AQI
Purpose to show AQI over time by city or compare cities

Density Map
Purpose to show visually AQI based on density by city



Average Monthly AQI
Purpose to determine if AQI varies based on time of year



OPEN - ANYTHING FROM MACHINE LEARNING/USER INPUT CITY?

Tool for Visualization/Dashboard: Tableau

Interactive element:

Future Draft Slides for Final Submission

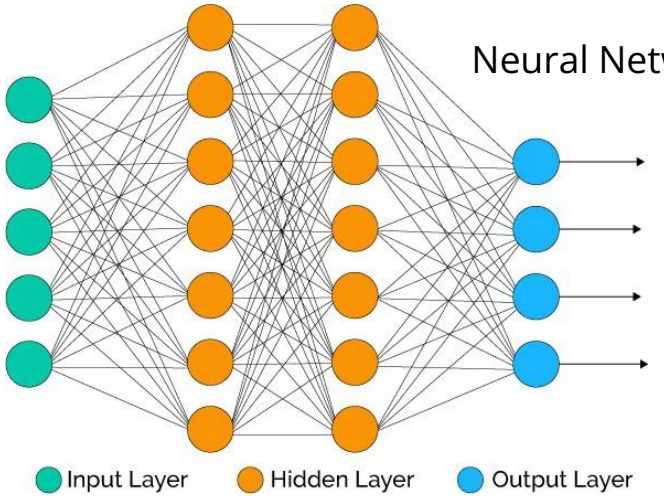
Technologies Utilized

Analyzing/Cleaning Data

Database

Technologies Utilized

Machine Learning Model



Dashboard

Other