Deliverable 1 Checklist: PR with code + report

1. **A brief introduction to your problem statement: DONE**
2. **Details of the data collection or cleaning steps you've undertaken**: DONE
3. Exploratory Data Analysis (EDA).
4. If your analysis has led to answers for any of the questions or if you've formulated hypotheses, especially for at least questions.
5. Individual contributions of each team member. We recommend that each team member writes 3-4 lines about their contributions, which can then be compiled into the report.

Problem Statement: Examine the influence transportation infrastructure has on the air quality and climate of Boston and surrounding neighborhoods.

- Poor air quality disproportionately affects communities in Boston, with the most adverse effects on marginalized communities. These impacts include increasing asthma rates and lung diseases.
- One possible solution to combating poor air quality involves transitioning public transportation to alternative energy sources, which can reduce harmful emissions and stabilize the air quality. However, this transition is rather complex.
- Since transportation plays an essential role in the lives of Boston residents, understanding the correlation between transportation infrastructure, air quality, and the health of residents can provide valuable insights into the necessary actions for Boston to implement.

Process Undertaken: Details of the data collection/cleaning

Data Collection

To conduct the analysis of transportation infrastructure and resulting air quality, a plethora of data is needed, some of which were provided to us by the City of Boston:

1. **Proximity to roads** (PPI Roads): Contains data about the spatial patterns of residents living in close proximity to roads with the highest levels of vehicle air pollution emissions across the MAPC region
    a. **g250mm_id**: Refers to a 250 x 250 mm area of population in Boston
    b. **nhwi_10**: Refers to the not-hispanic white population based on Census 2010 data
    c. **nhaa_10:** Refers to not-hispanic african-american population based on Census 2010 data
    d. **nhapi_10:** Refers to not-hispanic population asian-pacific islander population based on Census 2010 data
    e. **lat_10:** Refers to Latino population based on Census 2010 data
    f. **nhother_10:** Refers to other racial population based on Census 2010 data
    g. **ppi5:** Refers to air pollution emissions(**0: lowest, 5: highest**)

g250m_id	nhwhi_10	nhaa_10	nhapi_10	lat_10	nhoth_10	ppi5
144054	26.88	0.37	3.03	0.37	1.04	2
115030	33	0	14.59	0.31	2.1	1
232476	2.66	0	0	0	0	4
112471	1.34	0	0.04	0.02	0	0
148255	0	0	0	0	0	0
123090	13.99	1.12	2.48	0	1.8	1
107736	106.67	3.17	0.8	3.73	0.89	2

2. **Air Quality Data** (AQI Dashboard): Since the dashboard gets real-time data, use the AirnowAPI that feeds into the dashboard:AirNow API

Below is a screenshot of the output that is produced. For this insight, we have gathered AQI data for 2021.

Parameter	Description	Format	Example
dateObserved	Date of observation.	Date string (yyyy-mm-dd)	2012-02-01
hourObserved	Value of 0 for daily AQI.	Number	01
localTimeZone	Time zone of observed data value.	Text	PST
reportingArea	City or area name of observed data (data values are peak of monitoring sites associated with this area).	Text	Sacramento
stateCode	Two character state abbreviation.	Text	CA
latitude	Latitude in decimal degrees.	Number	38.33
longitude	Longitude in decimal degrees.	Number	-122.28
parameterName	Name of parameter.	Text	Ozone
aqi	Observed AQI value (peak value of monitoring sites associated with reporting area).	Number	45
categoryNumber	Observed AQI category number: 1. Good 2. Moderate 3. Unhealthy for Sensitive Groups 4. Unhealthy 5. Very Unhealthy 6. Hazardous 7. Unavailable	Number	2
categoryName	Observed AQI category name: • Good • Moderate • Unhealthy for Sensitive Groups • Unhealthy • Very Unhealthy • Hazardous • Unavailable	Text	Good

3. **Census(Transport, Income, Household Size) Data** (Census Bureau): In order to measure the AQI impact on population, it is necessary to get census data. The dataset we collected provided valuable insights into the modes of transportation, income, and households for 45 zipcodes in Boston for the year 2021. The image below shows the specific data recorded by the dataset.

```
Index(['Zipcode', 'Estimated_Civilian_Noninstitutionalized_Population',
       'Estimated_Civilian_Noninstitutionalized_Population_with_Health_Coverage',
       'Estimated_Civilian_Noninstitutionalized_Population_No_Health_Coverage',
       'Percent_Civilian_Noninstitutionalized_Population_with_Health_Coverage',
       'Percent_Civilian_Noninstitutionalized_Population_No_Health_Coverage',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Car, truck, or van -- drove alone',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Car, truck, or van -- carpooled',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Public transportation (excluding taxicab)',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Walked',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Other means',
       'Estimated_Total_Population_Commuting_to_Work_16_years_and_over_Worked from home',
       'Percent_Total_Population_Commuting_to_Work_16_years_and_over_Car, truck, or van -- drove alone',
       'Percent_Total_Population_Commuting_to_Work_16_years_and_over_Car, truck, or van -- carpooled',
       'Percent_Total_Population_Commuting_to_Work_16_years_and_over_Public transportation (excluding taxicab)',
       'Percent_Total_Population_Commuting_to_Work_16_years_and_over_Walked',
```

Data Cleaning:

1. **Census Data <LINK>**
    a. In order to have more organized data, we ran a Python script that dropped empty values and pre-processed data that was relevant for 2021.
    b. While the entire script can be found on the Github repository, here is a screenshot:

```
missing_percentage = DP02.isnull().mean() * 100
# Get columns with missing values
columns_with_missing = missing_percentage[missing_percentage > 0].index.tolist()
# Drop columns with missing values from the dataset
DP02 = DP02.drop(columns=columns_with_missing)
DP02['NAME'] = DP02['NAME'].str.replace('ZCTA5 ', '')
DP02.columns = DP02.iloc[0]
DP02 = DP02.drop(0).reset_index(drop=True)

specified_columns = [
    "Geographic Area Name",
    "Estimated Total Households",
    "Estimate!!HOUSEHOLDS BY TYPE!!Total households!!Average household size",
    "Estimate!!HOUSEHOLDS BY TYPE!!Total households!!Average family size",
    "Estimate!!PLACE OF BIRTH!!Total population",
    "Estimate!!PLACE OF BIRTH!!Total population!!Native",
    "Percent!!PLACE OF BIRTH!!Total population!!Native",
    "Percent!!PLACE OF BIRTH!!Total population!!Native!!Born in United States",
    "Estimate!!PLACE OF BIRTH!!Total population!!Native!!Born in United States",
    "Estimate!!PLACE OF BIRTH!!Total population!!Native!!Born in Puerto Rico, U.S. Island areas, or born abroad t
    "Percent!!PLACE OF BIRTH!!Total population!!Native!!Born in Puerto Rico, U.S. Island areas, or born abroad to
```

2. AQI Data <LINK>
    a. In order to have more organized data, we ran a Python script that converted the aqi.db files into csv for easy access via pandas.
    b. While the entire scripts can be found on the Github repository, here is a screenshot:

```python
# Adjusting the transformation function to also handle "PM10" and ignore unexpected parameters.
def extended_transform(row):
    # Parse the JSON data
    parsed_data = json.loads(row['data'])

    # Initialize a dictionary to store the transformed data with default values
    transformed = {
        'date': row['date'],
        'zip_code': row['zip_code'],
        'ReportingArea': '',
        'StateCode': '',
        'Latitude': 0,
        'Longitude': 0,
        'OZONEAQI': None,  # Default as None, which will be replaced by actual values or remain as NaN in the Dat
        'PM2.5AQI': None,  # Same as above
        'PM10AQI': None,  # Adding default for PM10
        'CategoryNumber': None,
        'CategoryName': ''
    }

    # Extracting information specific to "OZONE", "PM2.5", and "PM10"
    for record in parsed_data:
        parameter = record['ParameterName']

        if parameter not in ['OZONE', 'PM2.5', 'PM10']:
            continue  # If the parameter is not one of the expected types, skip it

        # Update the common information if not already done
        if not transformed['ReportingArea']:
            transformed.update({
                'ReportingArea': record['ReportingArea'],
                'StateCode': record['StateCode'],
                'Latitude': record['Latitude'],
                'Longitude': record['Longitude'],
                'CategoryNumber': record['Category']['Number'],
                'CategoryName': record['Category']['Name']
            })

        # Update the AQI values for the specific parameters
        transformed[f'{parameter}AQI'] = record['AQI']

    return transformed


# Load the new dataset
new_file_path = '../Datasets/AQI/2021_daily_aqi_data.csv'
new_data = pd.read_csv(new_file_path)

# Apply the transformation to each row in the new dataset
extended_transformed_data_2021 = new_data.apply(extended_transform, axis=1)

# Convert the results into a DataFrame
final_extended_df_2021 = pd.DataFrame(list(extended_transformed_data_2021))

# Saving the final DataFrame to a CSV file
final_extended_df_2021.to_csv('../Datasets/AQI/2021_Daily_Aqi_Data_Cleaned', index=False)

# Returning the shape of the final DataFrame and the first few rows to confirm the transformation
(final_extended_df_2021.shape, final_extended_df_2021.head())
```

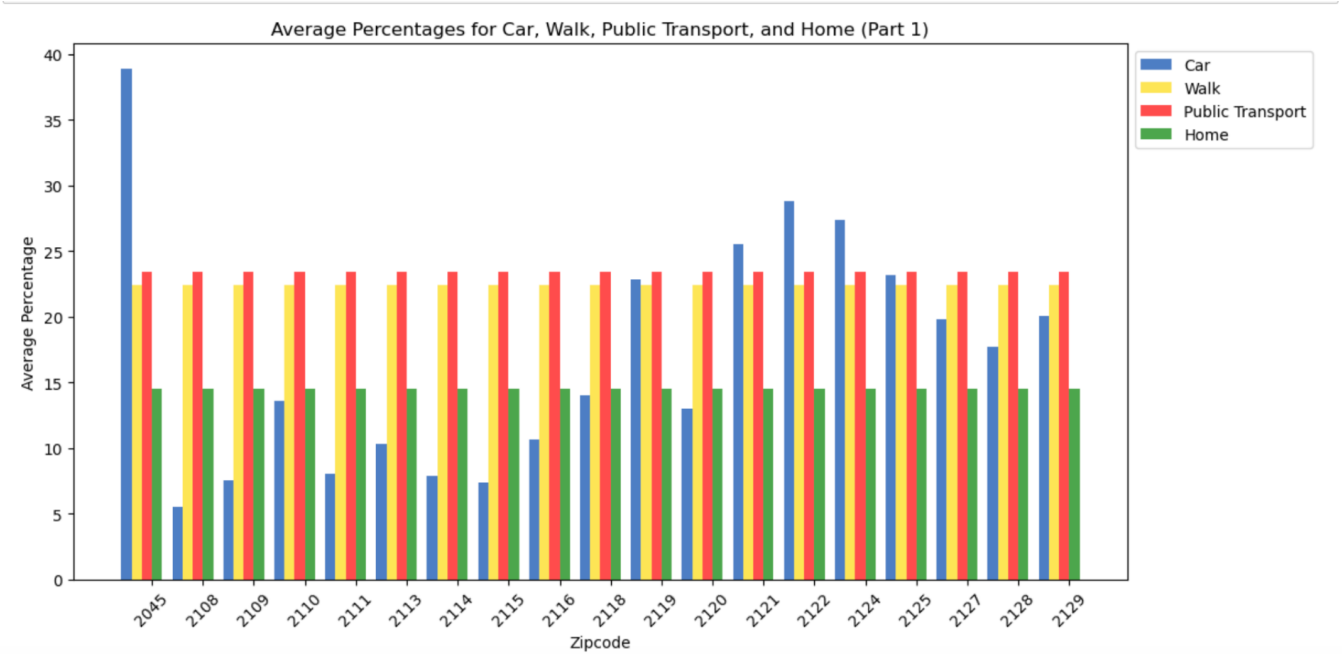
```python
path_file = '../Datasets/AQI/2021_Daily_Aqi_Data_Cleaned.csv'
df = pd.read_csv(path_file)

averages = df.groupby('zip_code')[['OZONEAQI', 'PM2.5AQI']].mean().reset_index()
averages['zip_code'] = averages['zip_code'].astype(str).str.zfill(5)

averages.to_csv('../Datasets/AQI/2021_Avg_Aqi_Data_Cleaned.csv', index=False)
```

First Insights:

Exploratory Data Analysis

- What is the yearly change in air quality for Boston residents based on their proximity to different types of transportation infrastructure specifically, proximity to public transportation options or proximity to roads?

Based on the grouped bar graphs below, we see that the Car seems to be the most popular medium of transport, followed by public transportation. What can be further studied is the impact this can have on air quality and how Boston combats this since the air quality for 2021 is relatively 'Good'. Although the overall air quality in Boston is relatively good, we hypothesize that as we look deeper into each zip code, the variance in air quality between zip codes will have some level of correlation to the relative distance to high emission areas (roads, highways, and public transportation lines).



How do areas with poor air quality compare to areas with better air quality based on different demographic characteristics, specifically:

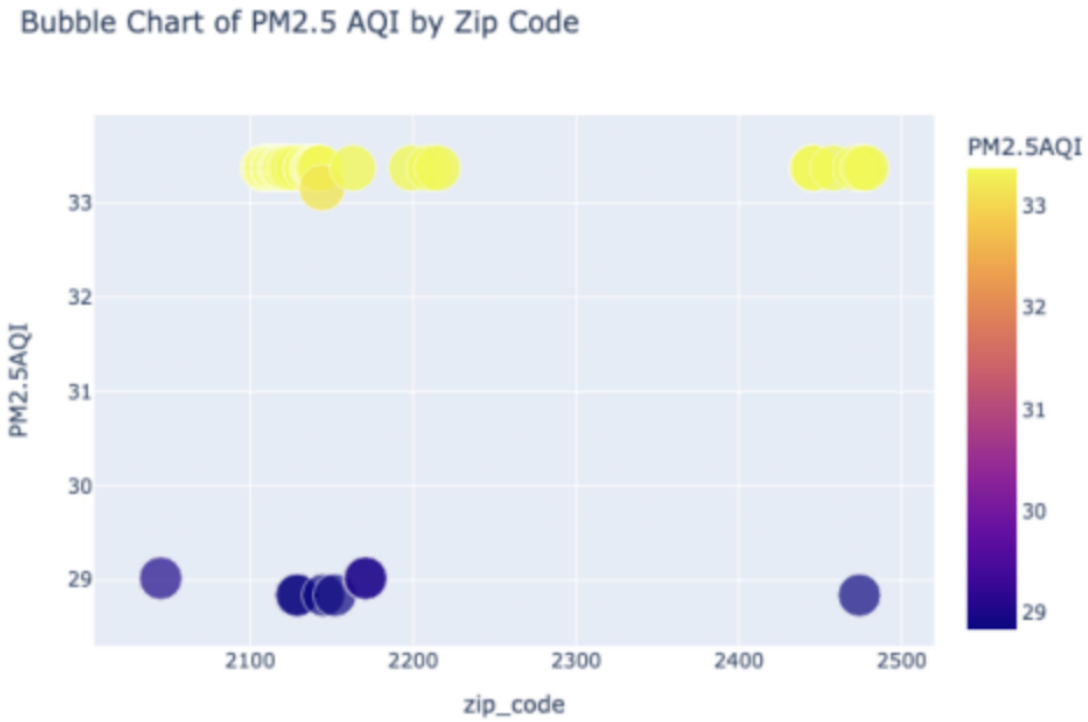
- Race/ethnicity (ACS)?

With respect to air quality and race impact, we noticed that as the PPI index decreased, the number of minorities decreased. On the other hand, as the PPI index increased, the number of White and African-American populations increased, with PPI index having the largest of these populations. We intend to look more into socioeconomic factors by race in each zip code, as this can give further insights into how disproportionate effects of air pollution may exist between white residents of Boston versus minority/POC residents. Factors like education history and median household income within a zip code/neighborhood can also relate to overall quality of life within that area (air quality, distance to green space, food deserts, public transportation).

- What are the trends in yearly change in air quality for Boston residents by neighborhood, zip code?

Based on the graph below, it is clear that for the year 2021, the yearly change in air quality was not too much since the average AQI stayed in the range of 31-32, indicating a generally good air quality for 96% of the 16,371 AQI readings throughout 2021. Although the overall air quality readings are 'good', we can see there are fluctuations in the air quality readings between roughly 29 - 33 within the same zip codes that we are interested in investigating to see what factors would create such variance.

	CategoryName	Count
0	Good	15699
1	Moderate	625
2	Unhealthy for Sensitive Groups	47



Next Possible Steps:

In order to give more depth answer to the first base question, the following can be steps can be taken:
Basically, the first base question asks the yearly change between air quality based on proximity to different types of transportation infrastructure. However, the given PPI dataset doesn't include useful information regarding the proximity. And further search for this information from different datasets seems not to yield any useful results too. So one way to deal with this is to estimate the proximity by the modes of transport from the Transport dataset. If this is done for each zip code, we would plot yearly change versus proximity. Eventually, we can create a linear regression of a polynomial regression model to answer this question.

Since the basic trends have been established, we can try to conclude about the yearly changes in Boston's AQI and the health impacts.

Conclusion:

In conclusion, we hope to dive deeper into each of the base questions as analysis continues. We ran into some roadblocks during our EDA with understanding the provided data and resorted to looking to outside sources to find similar data that was more comprehensive and required less cleaning. Now that we are aware of this, we definitely feel more prepared going forward to locate the data we need to perform deeper analyses and find new relationships between variables such zip codes, socioeconomic status, and proximity to high emission areas. Going forward, we want to look more into the MAPC data and understand how the $250m^2$ plots can be translated into zip codes so we can look into the ppi index and its relation to the racial breakdown of each area.

Individual Contributions:

Doruk Savasan:

- Fetched the Air Quality Data
- Fetched 4 different Census datasets (DP02, DP03, DP04, DP05)
- Cleaned the AQI as well as all the Census datasets
- Searched for correlations between AQI data and Census data

Medha Dhir:

- Fetched transport and zip code data.
- Explored the following trends:
- 1. AQI data for all zip codes
- 2. PPI Index Trends based on zip codes
- 3. Modes of transportation
- Worked on answering the base questions on the report

Maxwell Higa:

- I worked on decoding the PPI index data from the MAPC website, analyzing the data and creating some plots to show the relationship between each PPI index and the breakdown of population density for each racial group within each 250 square meter area. In deliverable 1, I worked on the conclusion, and added some of our hypotheses about the base questions and where to go from here.

Can Erozer:

- Analyzed the results from plots and determined what can be done next?
- Found solution to the problem that couldn't be solved directly with the given datasets
- Worked on finding extension project questions