Deliverable 1
DC City Council - SW Washington

Team 1:

1. Harshitha Tumkur Kailasa Murthy - U00683580
2. Vishwas Bhaktavatsala - U74206902
3. Kristopher Chou - U36116166
4. Yeh-Min Lu - U70603446

Sufficient data should have been collected to perform a preliminary analysis of the data and attempt to answer one question relevant to your project proposal which you will submit as a pull request.

1. **If data has already been collected for your project you must answer two questions.**

- We tried to explore different filters of US census data and were able to collect a few datasets according to the listed parameters and years.
- We collected data pertaining to age, sex, races and their populations in the counties present in the Southwest Washington Region.
- We have also collected data with respect to the education standards in these counties as well as the income details.

2. **Collect and pre-process a preliminary batch of data**

- We have collected a few datasets as mentioned above and the reason we chose these data sets are stated along with them.
- The collected dataset have a certain pattern to them, where they have data for each of the counties. Trying to map each county's data for all the datasets have been one of the objectives of our latest sprint.
- Using the Census API, we can pick out specific keys of data and make a table out of available data from District 11 in DP03 data. Doing so we can better visualize the data by comparing numbers to one another in a table. By using Jupyter Notebook to access the API, we are able to store certain data columns as variables and access them when we want.

3. **Perform a preliminary analysis of the data**

   We are looking to connect each of these datasets to understand how the income of families has affected the number of people in the house along with how educated the household would be. We would also like to understand the number of people who are educated and how the race has affected this number.

Answer 1-2 key questions

Few Analysis and related databases we have explored and downloaded datasets are for the below parameters:

Population demographics: We will analyze the distribution of the population in the Southwest cities in Washington DC based on race, sex, and age. This analysis will help us understand the overall demographic of the region and how it has changed over time. We can try to find out the patterns in the dense areas and how it is correlated with crime, education and poverty.

Age distribution: We want to analyze the age distribution in Southwest DC. This analysis will help us understand the age groups that are most represented in the region. We can try to find out if there are any patterns between areas with a higher or lower age average and how it correlates with crime, literacy rate, and poverty. We want to also analyze changes in the age distribution of the population over the past five years. This could involve examining trends in the proportion of elderly residents, changes in birth rates, and other factors that may be contributing to an aging population.

Racial and ethnic diversity: we want to analyze the racial and ethnicity of the population in the Southwest cities in Washington DC. This analysis will help us show the different groups that represent the region. We can try to find out the literacy rate and crime rates between different races and ethnicities and whether different groups have higher crime rates.

Gender diversity: we want to analyze the gender diversity of the population in the Southwest cities in Washington DC. This analysis will help us understand the distribution of males and females in the region. We can try to compare crimes caused by different genders.

Educational attainment: We want to analyze the educational attainment levels of the population in the Southwest cities in Washington DC. This analysis can help you understand the educational levels of the population in the region. We can analyse the data with men vs women literacy rate and how the household situation and wealth report has affected this metrics or data.

Income and poverty rates: We want to analyze the income and poverty rates of the population in the Southwest cities in Washington DC. This analysis can help you understand the economic conditions of the region as well as the well-being of their overall population.

Employment status and income inequality: We want to see the trends in employment status and income inequality over the past five years. This could involve examining trends in unemployment rates, changes in the proportion of workers in different industries, and changes in income inequality within the population.We are considering showing the different employment levels i.e, civilians, Armed forces ..,

Housing affordability: We want to analyze changes in housing affordability over the past five years. This could involve looking at changes in median home prices and rents, as well as changes in the proportion of residents who are paying a high percentage of their income on housing. Without affordable housing, families have constrained opportunities to increase earnings, causing slower GDP growth.Yeh0

Few Parameters we are majorly considering for demographical connection with various datasets:

1. Average income of women vs men
2. Death rate
3. Race and ethnicity in crime as well
4. Literacy rate

POC for which end tool to use for dashboard:

- ➔ Based on the initial POC on the tools to be used to visualize and dashboard the data would be best to use Tableau, Dundas BI or Google Charts.
- ➔ While all of them have their own pros and cons, Tableau has very good reviews regarding the visualizations it provides, its performance and ease of use.
- ➔ It has a very well written and exhaustive documentation and much required tutorial and training videos. Link.

POC for various ways to that the map visualization could be done:

- Shape files will be the best way to visualize the data as we can make an interactive map to hold our data
- Shape files allow us to create points, lines, vertex, and area features
    - https://library.carleton.ca/guides/help/shapefiles-help-guide (this is a good help guide)
- We will need to use ArcGIS to create shape files (it also uses python)
- We are able to use GIS files with .shp files because they can convert to one another
    - Common GIS files include:
        - AutoCAD Drawings
        - Digital Elevation Models
            - We can use these to map out the elevations of the city and compare to see if there are any differences in poverty levels, income, etc.
            - We can then use these models to add points using Shape files to outline certain patterns that appear.
        - Esri File Geodatabase
            - We can use these types of files to create heatmaps and visualize aspects such as income or demographics, creating a clearer visualization of all data laid out on a map.
        - Geography markup language
        - GeoJSON
        - LiDAR

POC for shape files visualization :

- A brief but detailed overview of what a shapefile is (video).
- A shapefile is a simple, nontopological format for storing the geometric location and attribute information of geographic features. Geographic features in a shapefile can be represented by points, lines, or polygons (areas). The workspace containing shapfiles may also contain dBASE table, which can store additional attributes that can be joined to a shapefile's features.
- Editing with shapefiles
    - We can edit shapefiles in ArcGIS Desktop with any license level (ArcGUS Desktop Basic, ArcGIS Desktop Standard, or ArcGIS Desktop Advanced). However, to take advantage of advanced editing functionality, such as topology, we will need to import the shapefile into a geodatabase as a feature class.
    - More above editing in ArcGIS.
- Importing shapfiles and dBASE tables to geodatabase feature classes and tables.
    - All feature types in shapefiles convert to geometry types in the geodatabase. Unlike coverages, shapefile feature types are similar to the geometry types stored in a geodatabase, so conversion is more straightforward.
    - How data converts when importing.
- Shapfile file size limitations
    - Each of the component files of a shapefile is limited to 2GB each, Therefore, .dbf files cannot exceed 2GB and .shp files cannot exceed 2GB (these are the only files that are likely to be huge). The total size for all the components files can exceed 2GB.
    - Creating a new shapefile.
        - A short video introduction into the creation and editing of shape files in arcmap and arccatalog.
    - Creating a new dBASE table.

Analysis on the Data:

- ➔ The Census data provides a vague image about the demographics of the South West Washington region. We tried to find the data based on the number of counties which are present in SW Washington.
- ➔ Based on initial analysis of all the data and datasets, it would require some effort to clean the data as well as find the required correlations.
- ➔ The Crime data is comprehensive and has the required latitude and longitude data, we can use this to point out and find patterns about crimes in regions of Washington.

Possible Quantitative Analysis

A quantitative data point for census data population demographics could be any numerical measurement or statistic that describes a particular aspect of the population being studied. For example, some common quantitative data points for census data could include:

Total population count: This would be the total number of people living in a particular geographic area, as determined by the census.
Age distribution: This would be a breakdown of the population by age range, such as the percentage of people who are under 18, between 18 and 25 and so on.
Gender distribution: This would be a breakdown of the population by gender, such as the percentage of people who identify as male or female.
Ethnicity or race distribution: This would be a breakdown of the population by ethnicity or race, such as the percentage of people who identify as white, black, Asian, Hispanic/Latino and Racial wealth report data source will help us do this.
Educational attainment: This would be a breakdown of the population by educational level, such as the percentage of people who have completed high school, college, a bachelor's degree, etc.
Income distribution: This would be a breakdown of the population by income level

Each of these data points can provide valuable insights.

Make sure to identify which team member is doing which tasks on the scrum report

Deliverable 1 tasks that are completed and In Progress:

Harshitha Tumkur Kailasa Murthy -

- Analysis on the datasets onto to which datasets would contribute towards the end goal - Done
- Understanding different data source blocks - Done
- Learnt and explored US census data and Racial wealth Gap data - Done
- Downloaded different datasets - Done
- Finding Datasets - found a few according to the earlier priority - Task In-progress
- Will work on Data Cleaning - Deliverable 2
- Preparation of Scrum report initial weeks - Done
- POC for which end tool to use for dashboard

Vishwas Bhaktavatsala

- Analysis on the datasets onto to which datasets would contribute towards the end goal - Done
- Understanding different data source blocks - Done
- Learnt and explored US census data and CDC data - Done
- Finding Datasets - found a few according to the earlier priority - Task In-progress
- Working on Data analysis - Deliverable 2
- Preparation of Scrum report last week
- POC for which end tool to use for dashboard

Kristopher Chou

- Analysis on the datasets onto to which datasets would contribute towards the end goal - Done

- Understanding different data source blocks - Done
- Finding Datasets found a few according to the earlier priority - Task In-progress
- Learnt and explored US census data and Open Data Catalog - Done
- Working on finding relationships and correlation in data and learning out how GIS files can be be created and used - task in-progress - Deliverable 2
- Preparation of Scrum report this week- done
- POC on how shapes files could be used
    - We can utilize all of the GIS files to create a big visualization of Southwest Washington DC, allowing users to click onto each aspect of the map and understand the pay discrepancies, if there are any, and also highlights patterns that are more prone to poverty and violence

Yeh-Min Lu

- Analysis on the datasets onto to which datasets would contribute towards the end goal - Done
- Understanding different data source blocks - Done
- Finding Datasets - found a few according to the earlier priority - Task In-progress
- Learnt and explored US census data and City Health Dashboard- Done
- Working on the implementation of geographic features in a shapfile with the representation of dBASE table. Task in-progress - Deliverable 2
- Preparation of scrum report next week
- POC on how shapes files could be used

Focus for Next Deliverable

1. Public safety and economic vitality - Rank 1
2. PS - effects of crime in southwest on different income categories
3. Gun violence and gun license record datasets to be looked into