Data Appendix:

CPI.csv

- This is a csv data file of the cpi values for each month from 1947 till 2024
- Unit of Observation: CPI value for each month
- The columns are 'DATE' and 'CPI'
    - 'DATE' is every month of every year from 1947 till 2024 in MM/DD/YY format
    - 'CPI' is the CPI value ranging from 20+ to 300+
- This file is cleaned to adjust house sale prices for inflation to 2024 dollars

CPI_clean

- This is the cleaned data file of the cpi values
- Unit of Observation: CPI average value for each year
- The columns are 'SaleDateYr' and 'CPI_Avg'
    - 'SaleDateYr' is the year of the CPI value in YYYY format
    - 'CPI_Avg' is the average CPI values from using every month in that year to calculate, value ranging from 20+ to 300+
- This file is cleaned to adjust house sale prices for inflation to 2024 dollars

Real_Estate_(Residential_Details)

- This is a raw file pulled from https://opendata.charlottesville.org/datasets that is used to get information about the houses in Charlottesville
- Unit of Observation: Houses in charlottesville
- The columns in this data are listed below, but only some are used in our analysis

RecordID_Int	ParcelNumber	StreetNumber	StreetName	Unit	UseCode	Style	Grade	Roof	Flooring	Heating	Fireplace	Year Built	TotalRooms	Bedrooms	HalfBathrooms	FullBathrooms	BasementGarage	Basement	FinishedBasement	Basement Type	External Walls	NumberOfStories	SquareFootageFinishedLiving

    - 'ParcelNumber' is the unique ID of each house
    - 'Grade' represents the quality of the house ranked by Charlottesviile in a letter grade
    - 'YearBuilt' is the year the house was built
    - 'SquareFootageFinishedLiving' represents the size in feet of the house
    - 'Bedrooms' is the number of bedrooms in the house
- This file is merged with the Real_Estate_(Sales).csv based on the 'ParcelNumber'

Real_Estate_(Sales).csv

- This is a raw file pulled from https://opendata.charlottesville.org/datasets that is used to get information about the sales of houses in Charlottesville
- Unit of Observation: Sales of houses in Charlottesville
- The columns in this data are listed below, but only some are used in our analysis

RecordID_Int	ParcelNumber	SaleDate	SaleAmount	StreetName	StreetNumber	Unit	BookPage

    - 'ParcelNumber' is the unique ID of each house
    - 'SaleDate' represents the date the house was sold in the format of YYYY/MM/DD HH:MM:SS
    - 'SaleAmount' is the amount the house was sold for

- This file is merged with the Real_Estate_(Residential_Details) based on the 'ParcelNumber'

merged_df_clean

- This is a merged file of the house sales from merging the Real_Estate_(Residential_Details) and Real_Estate_(Sales) based on the ParcelNumber where the values are cleaned. The CPI values are merged into the data frame too to adjust the SaleAmount values.
- Unit of Observation: Sales of houses in Charlottesville
- The columns in this dataset are:iParcelNumber, SaleAmount, SaleDate, Bedrooms, YearBuilt, SquareFootageFinishedLiving, Grade , SaleDateYr, Grade out of 10, CPI_Avg,
    - 'ParcelNumber' is the unique ID of each house
    - 'SaleAmount' is the amount the house was sold for
    - 'YearBuilt' is the year the house was built
    - 'SquareFootageFinishedLiving' represents the size in feet of the house
    - 'Bedrooms' is the number of bedrooms in the house
    - 'Grade' represents the quality of the house ranked by Charlottesviile in a letter grade
    - 'Grade out of 10' represents the quality of the house ranked by Charlottesviile adjusted on a numeric scale of 0 to 10
    - 'SaleDateYr' is the year the house was sold on YYYY format
    - Grade out of 10, CPI_Avg, SaleAmount_2024
    - 'CPI_Avg' is the average CPI values from using every month in that year to calculate, value ranging from 20+ to 300+
    - 'SaleAmount_2024' adjusted house sale prices for inflation to 2024 dollars
- This was used for our analysis

Uva Enrollment Data

- This is the data frame is used to be merged by the year with the 'merged_clean_df' to get the total enrollment for each year
- Unit of Observation: Year of Enrollment
- The columns in this dataset are:

Year	Total Enrollment	Undergraduate Enrollment	Graduate Enrollment

    - 'Year' is the year of enrollment in YYYY format
    - 'Total Enrollment' is the total enrollment per year
    - 'Undergraduate Enrollment' is the total undergraduate enrollment per year
    - 'Graduate Enrollment' is the total graduate enrollment per year

reg_df

- This is the data frame used to run our first set of testing
- Unit of Observation: Sales of houses in Charlottesville
- The columns in this dataset are:

SaleAmount_2024	SaleDateYr	Bedrooms	YearBuilt	SquareFootageFinishedLiving	Grade out of 10	TotalEnrollment

- ‘SaleAmount_2024’ adjusted house sale prices for inflation to 2024 dollars
- ‘SaleDateYr’ is the year the house was sold on YYYY format
- ‘YearBuilt’ is the year the house was built
- ‘Bedrooms’ is the number of bedrooms in the house
- ‘Grade out of 10’ represents the quality of the house ranked by Charlottesville
- ‘SquareFootageFinishedLiving’ represents the size in feet of the house
- ‘TotalEnrollment’ is the total graduate enrollment per year

- This was used to create a model to predict the SalesAmount

reg_df2

- This is the data frame used to run testing to understand the impact of just the TotalEnrollment
- Unit of Observation: Sales of houses in Charlottesville
- The columns in this dataset are:

SaleAmount_2024	TotalEnrollment

- ‘SaleAmount_2024’ adjusted house sale prices for inflation to 2024 dollars
- ‘TotalEnrollment’ is the total graduate enrollment per year

- This was used to create a model to predict the SalesAmount with only the TotalEnrollment to see the impact of UVA on Charlottesville