October 15, 2022

C964: Computer Science Capstone Template

Task 2 parts A, B, C, and D

[Part A: Project Proposal for Business Executives 3](#_Toc1515453419)

[Letter of Transmittal 3](#_Toc1044087116)

[Project Recommendation 4](#_Toc1721750458)

[Problem Summary 4](#_Toc205640471)

[Application Benefits 4](#_Toc739883765)

[Application Description 4](#_Toc1967641075)

[Data Description 4](#_Toc1580263866)

[Objectives and Hypothesis 5](#_Toc1084915425)

[Methodology 5](#_Toc1455093807)

[Funding Requirements 6](#_Toc1716359054)

[Data Precautions 6](#_Toc821607173)

[Developer’s Expertise 6](#_Toc742297278)

[Part B: Project Proposal 7](#_Toc1516064891)

[Problem Statement 7](#_Toc602847551)

[Customer Summary 7](#_Toc411678477)

[Existing System Analysis 7](#_Toc1668713332)

[Data 8](#_Toc1911399030)

[Project Methodology 8](#_Toc1307313575)

[Project Outcomes 8](#_Toc1967172832)

[Implementation Plan 10](#_Toc306674583)

[Evaluation Plan 10](#_Toc849271423)

[Resources and Costs 11](#_Toc1695791095)

[Timeline and Milestones 12](#_Toc787992355)

[Part C: Application 13](#_Toc1767968421)

[Part D: Post-implementation Report 14](#_Toc22579049)

[A Business (or Organization) Vision 14](#_Toc1496854501)

[Datasets 15](#_Toc645781746)

[Data Product Code 15](#_Toc157285634)

[Objective (or Hypothesis) Verification 17](#_Toc1315546441)

[Effective Visualization and Reporting 17](#_Toc656892821)

[Accuracy Analysis 19](#_Toc1956352798)

[Application Testing 20](#_Toc871280264)

[Application Files 20](#_Toc1639588139)

[User Guide 21](#_Toc1125023771)

[Summation of Learning Experience 21](#_Toc1702287338)

# Part A: Project Proposal for Business Executives

## Letter of Transmittal

September 18, 2023

Birkshire Hathaway Executives

Berkshire Hathaway Inc.

355 Farnam Street

Omaha, NE 68131

Dear Executives,

In recent years the housing market has made a drastic shift. Houses being listed are being sold at an alarmingly quick pace. Many people searching for homes to purchase are having a difficult time during the process due to this. As a real estate company, our focus should be to simplify and smooth out the process for organizations and individual’s using us to list properties as well as customers looking to use this company to purchase properties. Being able to quickly and accurately assess the value of a property will provide the greatest path to realize this goal and provide our customers the information they need when making these important decisions.

I am proposing a project that will use localized sales data to create a model that can evaluate potential listings and return a price that the house should be listed for based on properties with similar attributes. This project will allow us to maintain our status as a top notch real estate company in the market by allowing potential listings to be evaluated quickly and accurately.

This particular project will be developed with the ability for each branch of the company to utilize a set of data that is specifically tailored to the sales occurring in that branches immediate area. My computer science background will allow me to complete the programming for the project completely on my own, making development costs equivalent to my salary for the duration of the project. At my hourly rate of $40 an hour, development is estimated to take a total of 3 hours totaling $120. However, for each branch that would like to utilize the program, gathering data specific to that branch will require the bulk of expenditure for the project. A single data scientist should be able to comb through 2 years of sales data and coalesce it into a data set for the program in a two week period. At $20 dollars an hour for a total of 80 hours that comes to a total of $1600 for each branch that wishes to utilize the program. Furthermore, on a bi-annual basis each branch should update the data set for their location. This should take a data scientist roughly 3 shifts to update 6 months of new records and remove 6 months of the oldest records. This comes to a total cost yearly for each branch in maintenance $960.

Thank you for your time and I look forward to any questions or insights you may have during the development of this project.

Sincerely,

Joshua Gibson

Project Manager

[Jgib194@wgu.edu](mailto:Jgib194@wgu.edu)

## Project Recommendation

### Problem Summary

The project will develop a program that take input comprised of attributes for a potential property listing and then generate a listing price based on properties that have sold with similar characteristics. With the rapid pace of listings being sold in today’s market it is paramount for the company to be able to list new properties at a competitive price point within the market. It is even more paramount for the company to be able to evaluate potential listings as quickly and accurately in order for the company to maintain it’s competitive edge and serve our customers to our best ability. This project will deliver a program that can take input concerning a properties characteristics like number of rooms, size of the property, and age of the property and then provide a listing price based on properties with similar characteristics.

### Application Benefits

In order to properly serve this company’s consumer base the company needs to be able to accurately assess the value of a potential property listing. This program will allow that by utilizing data for sales that have already occurred in the real world. The program will use a regression model to map specific key features of a previous sale to the price point of the final sale. This will allow the organization to accurately assess a potential new listing and provide and accurate evaluation of the property value based on property sales that have occurred with similar features to the potential listing. Implementing the proposed solution will provide a more automated and impartial basis for potential listing evaluation. The solution will rely solely on facts and be able to generate an evaluation immediately. This will increase customer satisfaction by allowing sales agents to quickly assess a property, ascertain its potential value, and subsequently list the property at that value in a streamlined and effortless process, thus putting the property on display to potential purchasers with limited lag.

### Application Description

The application, for testing purposes, uses a data set that was provided from kaggle.com, however, the data set can be fine tuned for any particular area as needed. The data set utilized contains key features of a property like number of rooms, square meters of the property, distance from the city center, price, etc. The application takes this data as input and then utilizes a multiple linear regression model to map key features of a property (all features except price) to the price of sale for the property. The model then creates an algorithm that can predict the value of a potential property based on the key features of the property. A user will be prompted with a series of questions concerning the potential property. When each question has been answered, the program will utilize the inputs with the model to generate a price that the property should be sold as based on properties that have similar features.

### Data Description

The original data set was provided from kaggle at [Housing Market & Prices (kaggle.com)](https://www.kaggle.com/datasets/rukenmissonnier/final-house). Each data entry in the set is a unique collection of features for properties that have been included in the data set. There is a designation number for the entry followed by a set of features like number of rooms, square meters, distance from the city center, etc. ending in price. All of these features of a property are quantitative aspects. Price is the dependent variable in this scenario; all other features contained withing the data set are independent variables. When the program takes input based on these independent variable the model will then calculate an appropriate price evaluation, making price the definitive dependent variable in this scenario. Observations based on scatter plots of the data in multiple configurations has shown that there are virtually no anomalies to the data set currently utilized. These scatter plots are provided directly on launch of the program to provide the user a baseline to understand the data that was used for the underlying algorithm.

### Objectives and Hypothesis

This project hypothesizes that when given a set of input responses concerning the features of a property, a multiple linear regression model can evaluate those features and generate an accurate assessment of the properties value based on sales of properties with similar features. The desired outcome for the project is an accurate assessment of a potential property’s value. R-Squared will be utilized as a testing metric to verify the accuracy of the projects predictions versus a test set of data that the model has never encountered before. An R-Squared value of 0.9 is desired for accuracy.

### Methodology

For this project the incremental methodology will be utilized through all stages of development. This methodology will allow for the greatest collaboration between the development team and the customers this project aims to help. It will also allow for teams to be more flexible during the development stages of the project. The following five stages will occur during the project development:

**Ideation**:

During this stage the stakeholders of the project will gather and discuss the desired outcome for the project. Stakeholders will have the ability to bring forth questions and concerns about the project as well as provide more insight into the desired outcome of the project. These outcomes will be document as business and user requirements for project completion. The project will be broken down into various task that will then be distributed among the development team.

**Development**:

The development team will begin working on the project during this stage. As tasks are completed the development team will showcase the completed tasks to stakeholders as needed for additional feedback. Change requests will be handled by the change advisory board.

**Testing and Quality Assurance:**

The sample data will separated into two categories by the program. The first set will contain data entries that will be used for the training of the Multiple Linear Regression model being utilized for the project. This set will be known as the Training Set. The second set will contain a subsection of the original data set and will be utilized to test the model for accuracy. This set will be known as the Test Set. Once the Multiple Linear Regression model has processed the training set and established it’s baseline the test set will be introduced to the model and the R-Squared value will be calculated to establish the accuracy of the models predictions.

**Deployment**:

After thorough testing, the finished project will be released to the company’s branches for use. However, each branch has a different demographic when it comes to property values base on it’s local circumstances and sales records. Each branch that wishes to utilize the program would need to compile a data set of local sales that would be added to the program to be utilized as the training data set in order for results to be fine tuned to that branches local circumstances. From the programming perspective this is easily achieved by compiling a .csv file for each branch containing their local information and when the program is installed at each branch their local .csv will be used in the program.

**Maintenance**:

The housing market across the country shifts regularly. In order for the program to run optimally for each branch, periodic reviews of recent sales local to the branch will need to be conducted. At least on a bi-annual basis. During this review period the .csv file will be updated with recent sales data and the oldest entries equaling the time frame being updated will be removed. This will provide each branch with a robust data set that is relevant and recent to each branches demographic.

### Funding Requirements

The project will utilize the PyCharm IDE as well as various open source and free libraries through all stages of development and deployment of the project. Thus, there will be zero cost requirements for development beyond my hourly rate of $40 for three hours at a total of $120 to development the program. The main costs are in the data sets required for each branch of the company to utilize the program optimally. Each branch must gather a data set of local property sales in order for that branch to utilize the program effectively. A data scientist at each branch could gather two years of sales data throughout the course of two weeks. At an hourly rate of $20 per hours for a total of 80 hours this comes to a total of $960 for each branch to gather the appropriate data for their local demographic.

To provide the best customer satisfaction and to remain current with recent changes in the local housing market each branch will need to conduct a bi-annual review of market data. This will involve deleting six months of the oldest data for each branch as well as gathering the latest six months of the branches local sales data and incorporating it into the data set. A data scientist can perform this function every six months in three shifts. At six shifts total for the year at eight hours a day and $20 per hour this comes to a total of $960 per year for each branch to maintain the data set relevant to their local operations in order to provide customers with the most recent and up to date results that are desired.

### Data Precautions

The data utilized for the development of the project contains no sensitive or protected data. It is a data set obtained from kaggle.com that is free and open source. However, data utilized for each branches utilization of the project will be hand collected by a data scientist. This data set that will be collected will not contain any sensitive or protected data either. It will be gathered from public websites like zillow.com for the specific area of each branch.

### Developer’s Expertise

This project requires knowledge in machine learning as well as programming languages and programming methodologies to be completed. My degree in Computer Science from WGU has provided me with ample experience to fulfill all of the obligations. The project will

# Part B: Project Proposal

The project proposal should target your client’s technically savvy IT (Information Technology) professional leadership. Use appropriate industry jargon and sufficient technical details to describe the proposed project and its application. Remember, you’re establishing the technical context for your project and what it will accomplish for the client. Typically, this section is 8 – 10 pages. **Write everything in the future tense.**

## Problem Statement

There are a lot of factors that come into play when a home is sold. How many rooms? What is the square footage? How far is the property from certain amenities like the center of a close large city? All of these factors will go into play when evaluating a potential property and determining the properties value in the open market. A companies ability to do this quickly and accurately will make it or break it when coming to customer satisfaction and revenue generated for the company. Accurate assessments of a property and quicker sales will increase customer satisfaction and revenue generated by the company.

## Customer Summary

Homes. They are a very valuable asset and desired commodity across the world. The customers that will be targeted with this proposal are individuals that wish to sell their homes on the open market. Our customers are those individuals that wish to place their homes on the market for sale. But how do we decide the listing price for the property? In the past this would generally be based on the intuition of the listing agent. The proposed data product will take intuition out of the equation. Instead, property evaluations will be developed based on properties that contain similar characteristics of the potential listing.

## Existing System Analysis

Generally, for a person to sell their house they would contact a real estate agent; you could sell your home on your own as an option but not advisable due to legalities. The listing agent would then meet with you and take a look at the property and then give a guestimate to what the property should be listed at. The customer would sign off and then the property would be listed by the real estate agent.

The real estate agent would then conduct open houses for potential buyers. Potential buyers would tour the homes during these events to get a feel for the property. In some circumstances the potential buyers would deem the property overpriced for their taste or simply do not like the property. Obviously in these circumstances no sale will occur. In other circumstances potential buyers like the property and will put in an offer for purchase. Now this offer for purchase does not have to be the amount that the property was listed for. It can be lower or even higher based on the demand in the local market. Cash versus loan offers also factor into this. However, the crux of the matter when initially listing a property is what will it sell for, and currently that is a guessing game made by the listing agent hired by the seller.

I would like to divulge the experience of a purchaser that will highlight the current issue, my own personal journey to home ownership. All of the following events that I will outlay are typical steps involved during the sale and purchase of a house. The house was listed at the price of $189,000 before I even knew of its existence. After some time the listing price was changed to $179000 and it was a few weeks after this drop that I had noticed the property. I scheduled a visit, like the property and solicited my real estate agent to put in an offer for $169,000. After a day of back and forth they accepted an offer of $172,000. Once an offer is accepted there would be a property inspection to verify that everything they said about the property was true and there are no defects not listed like termites and such. Then, if you are utilizing a loan, the loan originator would require an appraisal of the property by their appraisal company. The interesting part of this from my perspective is that the appraisal company didn’t give an actual value the property was appraised at, instead documentation stated that the property was appraise for the value it was being sold. I highlight this because as we can see from this particular real world example the housing market currently is a bit wishy washy. Real estate agent “assesses” the property and would want it to list if for a certain price, then negotiations if it isn’t to purchaser’s ideals, then what would it be evaluated at by the insurer, and on and on. This solution will provide more accurate and up to date evaluations of properties based on recorded sales data for properties with similar features without subjective interpretation.

## Data

The original data set was provided from kaggle at [Housing Market & Prices (kaggle.com)](https://www.kaggle.com/datasets/rukenmissonnier/final-house). Each data entry in the set is a unique collection of features for properties that have been included in the data set. There is a designation number for the entry followed by a set of features like number of rooms, square meters, distance from the city center, etc. ending in price. All of these features of a property are quantitative aspects. Price is the dependent variable in this scenario; all other features contained withing the data set are independent variables. When the program takes input based on these independent variable the model will then calculate an appropriate price evaluation, making price the definitive dependent variable in this scenario. Observations based on scatter plots of the data in multiple configurations has shown that there are virtually no anomalies to the data set currently utilized. These scatter plots are provided directly on launch of the program to provide the user a baseline to understand the data that was used for the underlying algorithm.

The ultimate goal of this project is to provide each branch of the company with the ability to determine property values of potential listings in their area based on their local area sales. For the design , development, and testing stages the sample data set from kaggle will be utilized. During the deployment stage a data scientist will be employed to gather data from each branches local area sales using zillow.com and coalescing that data into a branch specific csv file that can be incorporated into the program for the specified branches deployment. During the maintenance stage the data scientist associated with each branch will update the csv file for their designated branch every six months with the most recent six months of sales data for the area. They will also delete the oldest six months of data from the csv file. This will allow for the data set to contain a consistent two year range of data for the model to utilize. Incomplete and outlier data anomalies will be ignored by the program.

## Project Methodology

For this project the incremental methodology will be utilized through all stages of development. This methodology will allow for the greatest collaboration between the development team and the customers this project aims to help. It will also allow for teams to be more flexible during the development stages of the project. The following five stages will occur during the project development:

**Ideation**:

During this stage the stakeholders of the project will gather and discuss the desired outcome for the project. Stakeholders will have the ability to bring forth questions and concerns about the project as well as provide more insight into the desired outcome of the project. These outcomes will be document as business and user requirements for project completion. The project will be broken down into various task that will then be distributed among the development team.

**Development**:

The development team will begin working on the project during this stage. As tasks are completed the development team will showcase the completed tasks to stakeholders as needed for additional feedback. Change requests will be handled by the change advisory board.

**Testing and Quality Assurance:**

The sample data will separated into two categories by the program. The first set will contain data entries that will be used for the training of the Multiple Linear Regression model being utilized for the project. This set will be known as the Training Set. The second set will contain a subsection of the original data set and will be utilized to test the model for accuracy. This set will be known as the Test Set. Once the Multiple Linear Regression model has processed the training set and established it’s baseline the test set will be introduced to the model and the R-Squared value will be calculated to establish the accuracy of the models predictions.

**Deployment**:

After thorough testing, the finished project will be released to the company’s branches for use. However, each branch has a different demographic when it comes to property values base on it’s local circumstances and sales records. Each branch that wishes to utilize the program would need to compile a data set of local sales that would be added to the program to be utilized as the training data set in order for results to be fine tuned to that branches local circumstances. From the programming perspective this is easily achieved by compiling a .csv file for each branch containing their local information and when the program is installed at each branch their local .csv will be used in the program.

**Maintenance**:

The housing market across the country shifts regularly. In order for the program to run optimally for each branch, periodic reviews of recent sales local to the branch will need to be conducted. At least on a bi-annual basis. During this review period the .csv file will be updated with recent sales data and the oldest entries equaling the time frame being updated will be removed. This will provide each branch with a robust data set that is relevant and recent to each branches demographic.

## Project Outcomes

The project upon completion will contain the following items:

* The finished application
* A usere guide.

## Implementation Plan

The project will be implemented on a branch by branch basis. Phase one of the rollout will be launched with branches that have signed up for the initial roll out of the project and have approved the proposed costs for their branch. Considering that final deployment of the application concerns locally sourced data this will provide an opportunity for the development team to utilize additional testing using locally sourced data to verify the accuracy of the linear regression model that has been deployed for the project. Testing will be conducted thoroughly before deployment, but this rollout is heavily dependent on the local area sales data available to provide each branch with up to date local data. It is paramount that the local are data has been meticulously gathered and formatted to the original data set used for development and testing.

* Provide an outline of how the project will be implemented. This description might include the following:
  + General strategy.
  + Phases of the rollout.
  + Dependencies.
  + Details for testing and distribution.

## Evaluation Plan

**Ideation**:

The project has very specific and defined business objectives that have been outlined. For each desired element that will be incorporated within the project there will be an evaluation on whether that objective aligns with the business objectives. If it does, it will be included in the final product upon project completion. However, if it does not conform to any business objective it will not be incorporated into the project.

**Development**:

Through all stages of development regular checks will be conducted to verify that the production being conducted by the development team adheres to the desired results proposed by the ideation stage as well as the business objective outlined for the project.

**Testing and Quality Assurance:**

The sequestered testing data set will be introduced to the model within the program. The R-Squared value will be based on the models predicted price values versus the actual price values contained in the data set will be determined. A R-Squared value closer to one will show that the model can accurately predict a price evaluation of a potential property.

**Deployment**:

During this stage, each branch that wishes to utilize the new program will need to gather local sales data of properties in their area. Once that data has been gathered it will be introduced into the Multiple Linear Regression model. The data set, for testing purposes will be separated into a training set and a testing set similar to the development stage. The training set for each branch will be used to train the model for their individual branch and the testing set will be utilized for testing purposes. The same stragety utilized during the development stage will be utilized here…calculating the R-Squared of predictions versus actual data records. An R-Squared closer to one will indicate a success.

**Maintenance**:

Biannual gathering and deletion of six months worth of records for each branch will be conducted in order to keep each branch up to date with their local market conditions. This new data set will become the training and testing data sets utilized in other stages of development. The same process of calculating the R-Squared value of predictions from the model for the training set will be utilized to ensure that the model is up to date and providing accurate predictions to the users that employ the program.

## Resources and Costs

Each branch already has the required hardware for project implementation…a computer. The software development will be performed in house by myself and at my hourly rate of $40 per hour it will take three hours to complete the project for a total of $120 dollars for the project completion. The program that will be developed utilizes free and open source libraries, so there will be no additional costs for the implementation of the final program. No licensing fees will be associated with the project at any stage of development.

The brunt of the project’s cost comes from the fine tuning of the project to each branch that implements the finished program. Housing markets vary from one area to the next. Very large high cost of living cities will have a vastly different market then say smaller more rural communities. Therefore, each branch must acquire a data set that is current and relevant to the branches locale. In order for this to occur a data scientist is needed for each branch initially to aggregate local sale transactions into a CSV file that the program can utilize. This local CSV file will provide the core data when training and testing the program for utilization by the branch to provide the most robust and accurate model algorithms for the program to utilize when being used for it’s intended purpose. A data scientist at each branch could gather two years of sales data throughout the course of two weeks. At an hourly rate of $20 per hours for a total of 80 hours this comes to a total of $960 for each branch to gather the appropriate data for their local demographic.

I will perform the deployment stage myself. For each branch that utilizes the program it will take approximately one hour to implement the new program in each machine for the branch location and to couple the branches local CSV file with the branches program. Maintenance of the program requires the utilization of a data scientist every six months.To provide the best customer satisfaction and to remain current with recent changes in the local housing market each branch will need to conduct a bi-annual review of market data. This will involve deleting six months of the oldest data for each branch as well as gathering the latest six months of the branches local sales data and incorporating it into the data set. A data scientist can perform this function every six months in three shifts. At six shifts total for the year at eight hours a day and $20 per hour this comes to a total of $960 per year for each branch to maintain the data set relevant to their local operations in order to provide customers with the most recent and up to date results that are desired.

## Timeline and Milestones

* Provide a projected timeline, including start dates and end dates for each milestone (a table is acceptable).

|  |  |  |
| --- | --- | --- |
| **Stage** | **Start Date** | **Finish Date** |
| Ideation | 11/06/2023 | 11/12/2023 |
| Development | 11/13/2023 | 11/15/2023 |
| Testing | 11/16/20023 | 11/19/2023 |
| Deployment | 11/20/2023 | 12/3/2023 |
| Maintenance | Every 6 months | Every 6 months |

# Part C: Application

The following items will be submitted:

* Zip file containing the project
* User guide.

# Part D: Post-implementation Report

Create a post-implementation as outlined below. Provide sufficient detail so that a reader knowledgeable in computer science but unfamiliar with your project can understand what you have accomplished. Using examples and visualizations (including screenshots) beyond the three required is highly recommended. **Write everything in the past tense.**

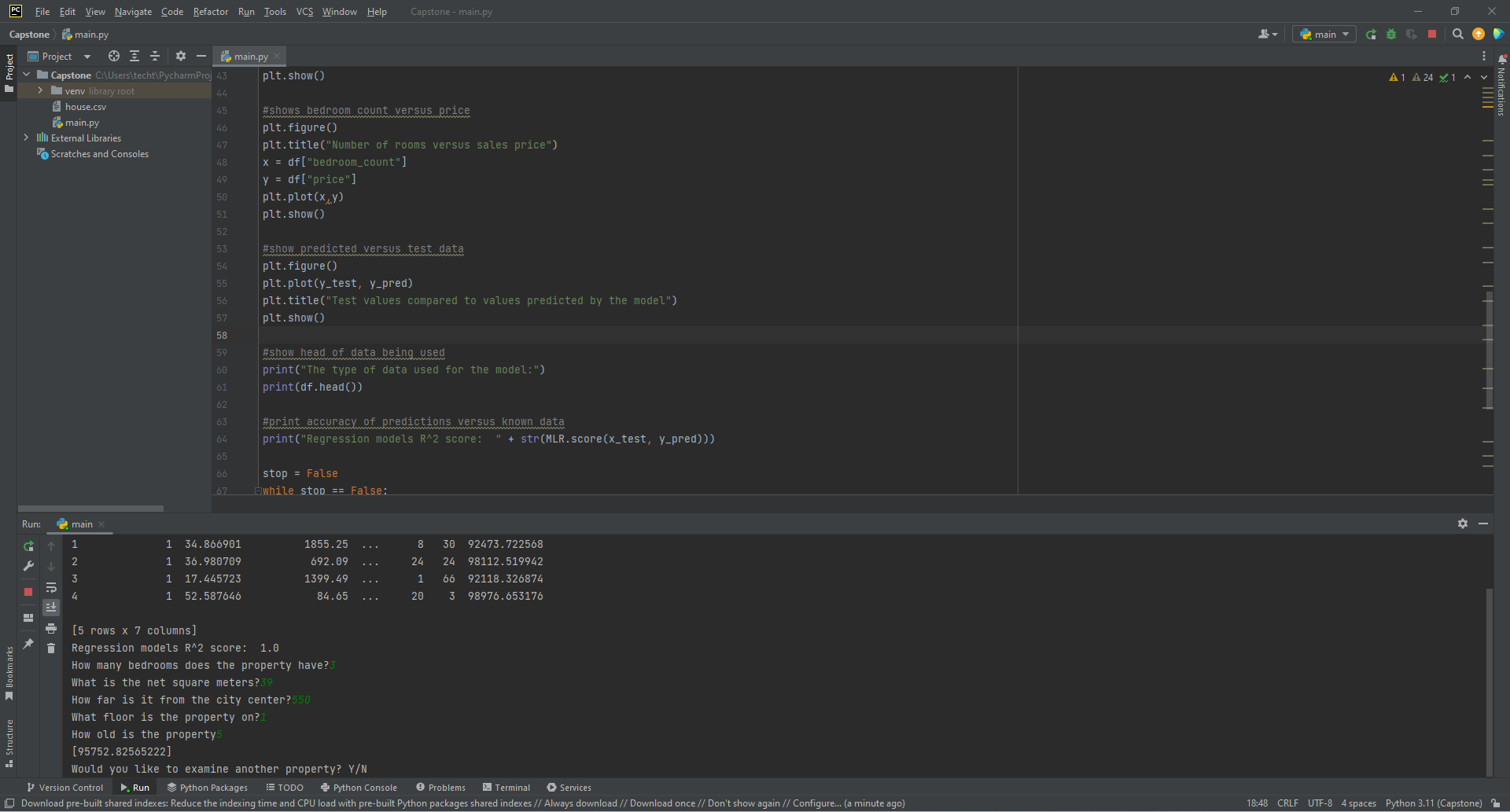
## A Business (or Organization) Vision

The housing market fluctuates regularly throughout the year. This fluctuation is affected by a variety of factors like prime interest rates as well as sales instances contained within a local area. Housing market evaluations, much like vehicles, are largely based on local market rates and the willingness for a customer to purchase at the suggested rate. Vehicle pricing is generally lower stakes negotiation, whereas a listed property can regularly receive potential offers that are lower than the asking price. So the problem becomes, how do we accurately evaluate the price of a potential propety that should be listed.

The current model for evaluating a property is for a customer to enlist a real estate agent in their area to facilitate a sale. This real estate agent will meet with the customer and evaluate the property. They will provide a guestimate evaluation of the property based on the properties features and their personal experience. It is listed, an offer is made, if accepted an appraisal company verifies if the property is worth the amount of the offer. The problem is that this evaluation is made by a human, who can make mistakes.

To address the current problem of potential human mistakes, this program uses a Multiple Linear Regression machine learning model to remove the human portion from the equation when determining an accurate evaluation of a potential property listing. The model takes characteristics of a property, known as features, as an input to create an algorithm that maps those features to final observed sales prices. Once this algorithm has been established a user can input a sequence or parameters that align with the categories of the original features to determine an evaluation of a property based on local sales data obtained by the company.

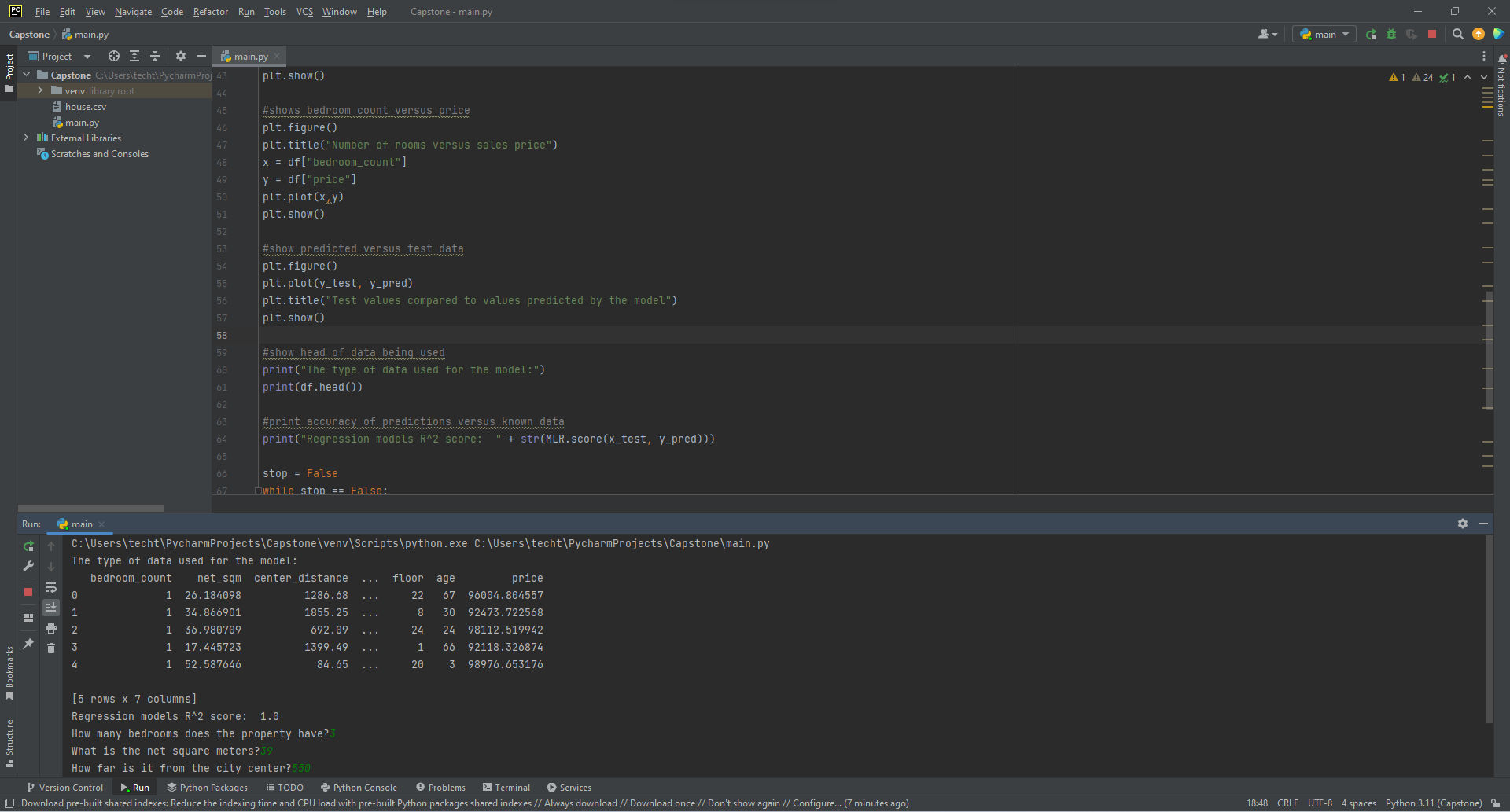
The following example shows how the user can interact with the program and determine a property evaluation:



## Datasets

For the purposes of development and testing the data was gathered from kaggle.com at [Housing Market & Prices (kaggle.com)](https://www.kaggle.com/datasets/rukenmissonnier/final-house). This data set was not processed because there were no outliers for the data set. During the deployment stage, each branch that utilizes the program collected local data sales from zillow.com. There was no need to process this data because the fields specific to the program were the only fields gathered during the data collection process.

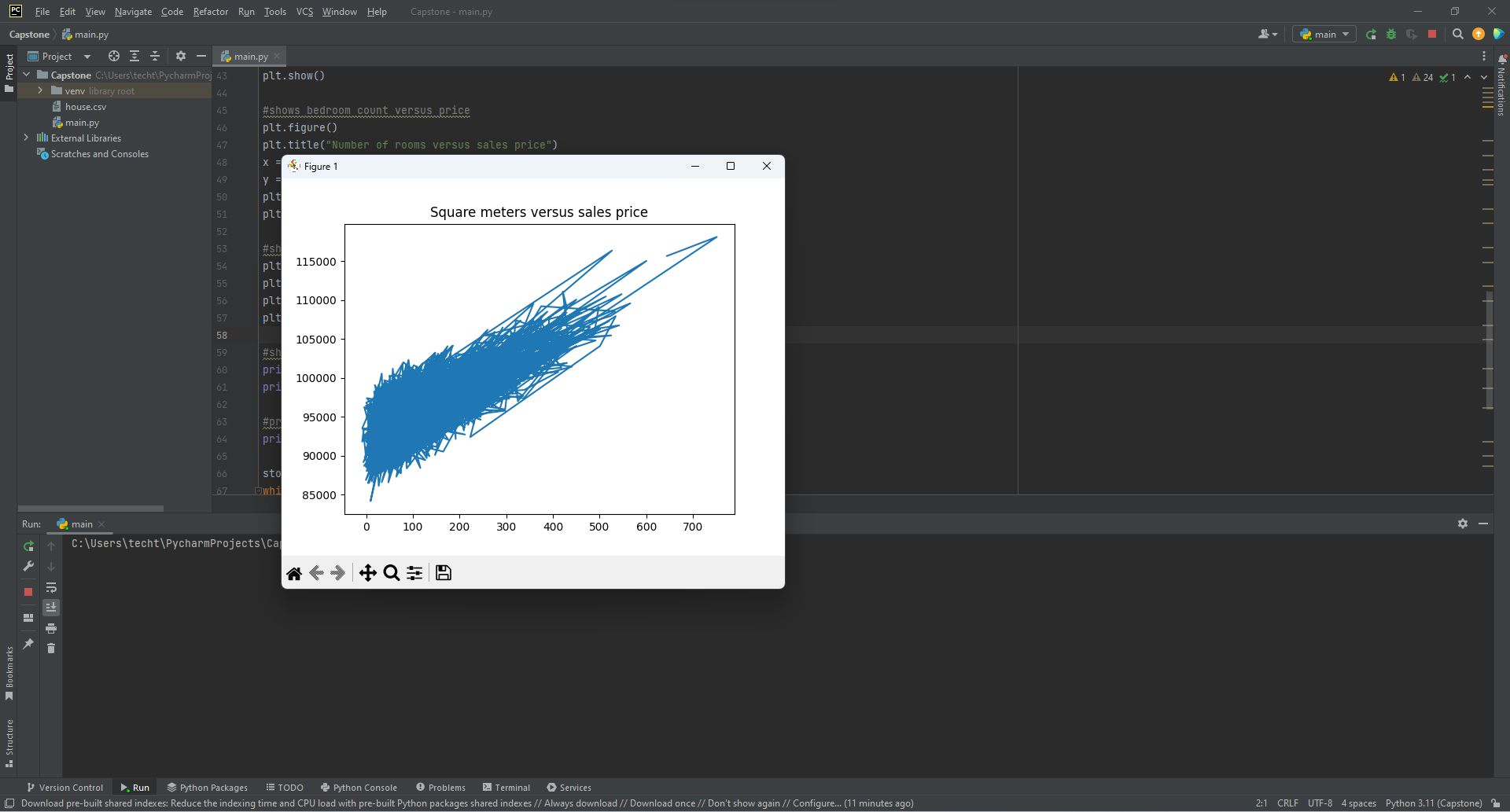
Data for both the development data and each branches specific data contain the same categories. The following depicts what categories were present in the data set.



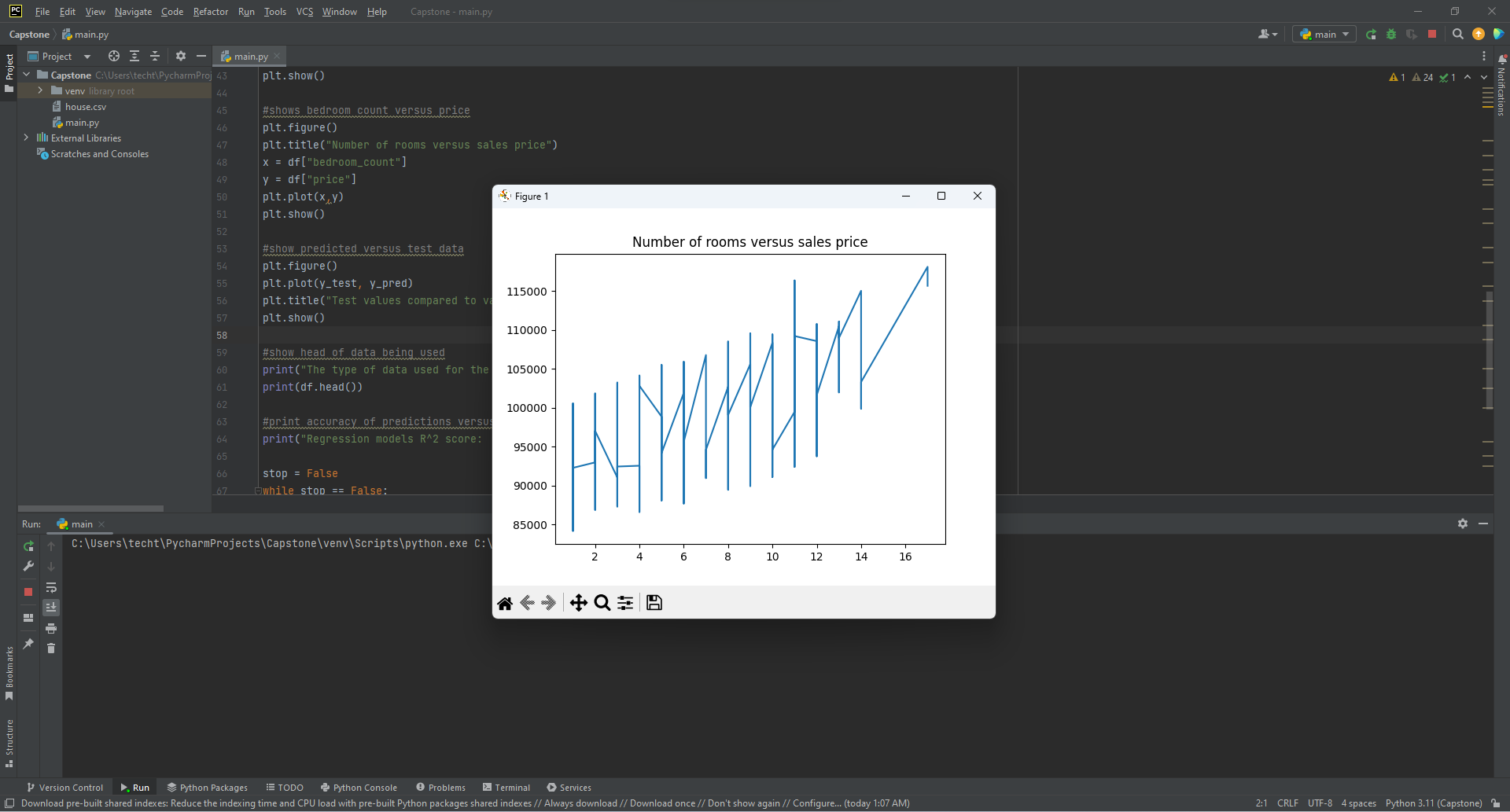
## Data Product Code

To better understand the data being used by the program the development team implemented a series of figures that would pop up in succession upon running the program. None of the data was processed because it provided all of the relevant information desired by the development team. These figures provided a descriptive visualization of the data being utilized. Three separate visualizations were provided by the developers.

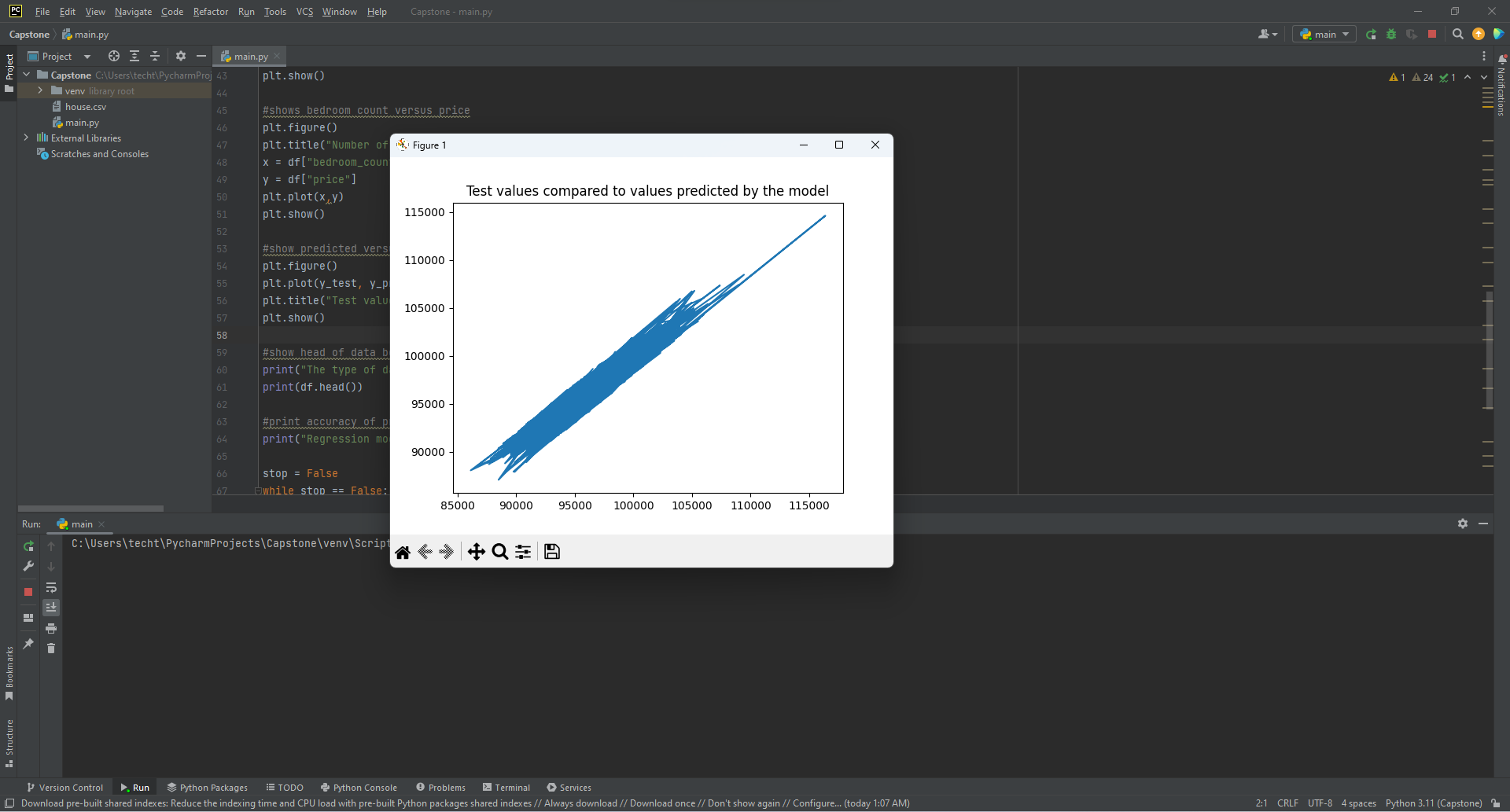
First there is a table showing the data in terms of the square meter of a property versus the sale price of the property:



The next provides a breakdown of the number of rooms associated with a property and its final sale price:



Finally, the program provides a visual analysis of the data being used to test the model. This visualization plots the predicted value from the model against the actual sale price contained within the test data site:



To achieve the predictive results desired a Multiple Linear Regression Model was utilized for the non descriptive portion of the project. This model takes a set of multiple features, in this case quantitative descriptors of a property and then generates a predicted sales price based on those features. The model itself was trained with a data set. During the development stage a data set was acquired that was segmented into to separate sets. During the segmentation a large portion was designated as the training set, roughly 70%, and the remaining was designated as the testing set. The training set was introduced into the model to develop the final algorithm that would be utilized to predict property values. The testing set was features were then introduced into the model and the predicted sale price by the model was compared to the price contained in the testing set to determine accuracy of the model.

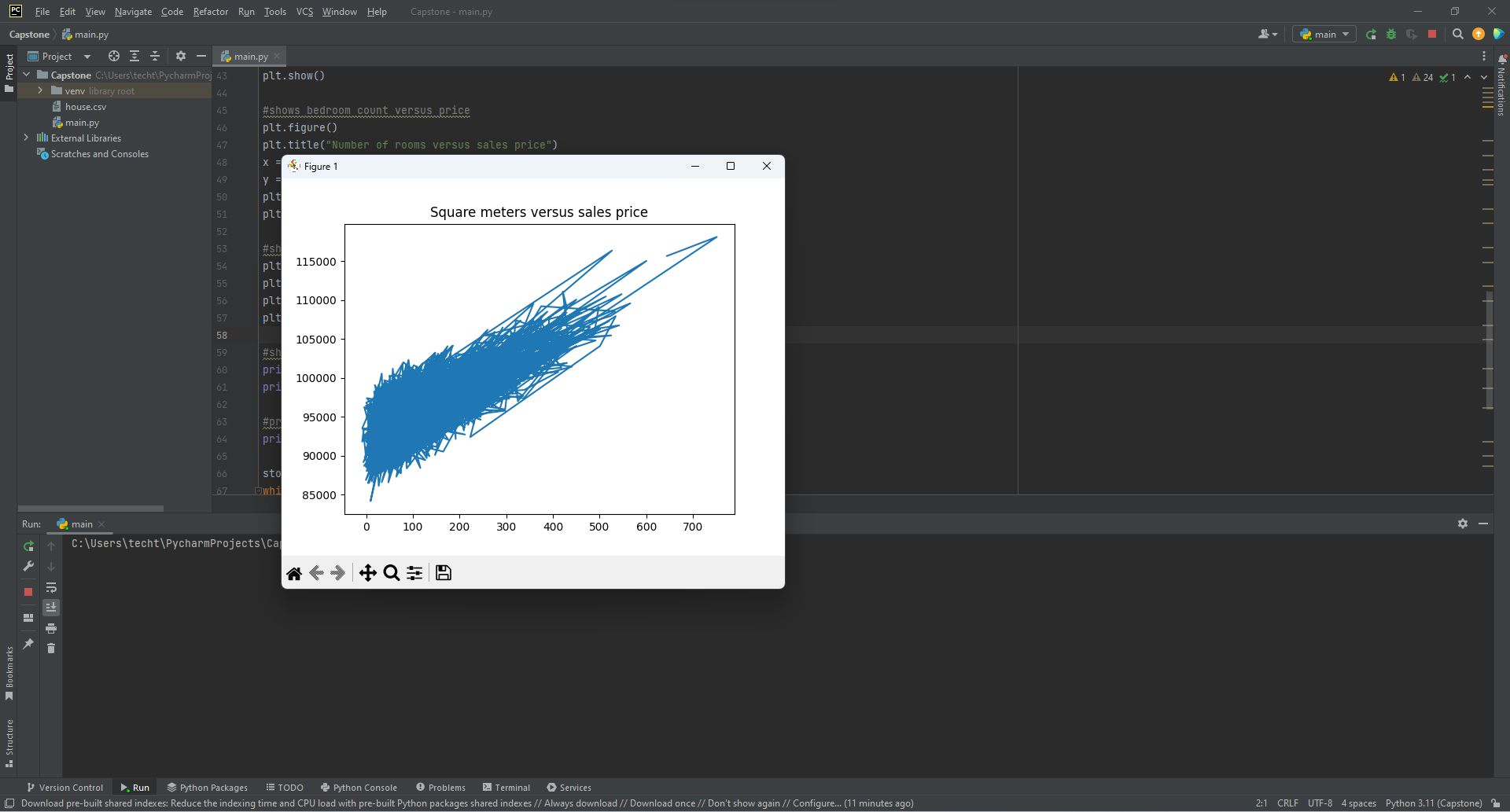
## Objective (or Hypothesis) Verification

The project hypothesizes that given a set of features that describe a home a prediction can be made to evaluate the sale price that the home should be sold as based on recently sold homes that have similar features of the potential property. The program utilizes a Multiple Linear Regression model to utilize a training set to calibrate the model. Once the calibration has occurred, a subset of the data that has never interacted with the model is introduced to the model to provide predictions based on the features that have been designated as core evaluation metrics for the model. The models predictions are then compared to the actual price values for that set of data and the accuracy of the model is calculated via a metric called R-Squared. According to the R-Squared calculations of the model versus actual data at a rate of 1.0 the predictions are highly accurate. Therefore, the objective has been met.

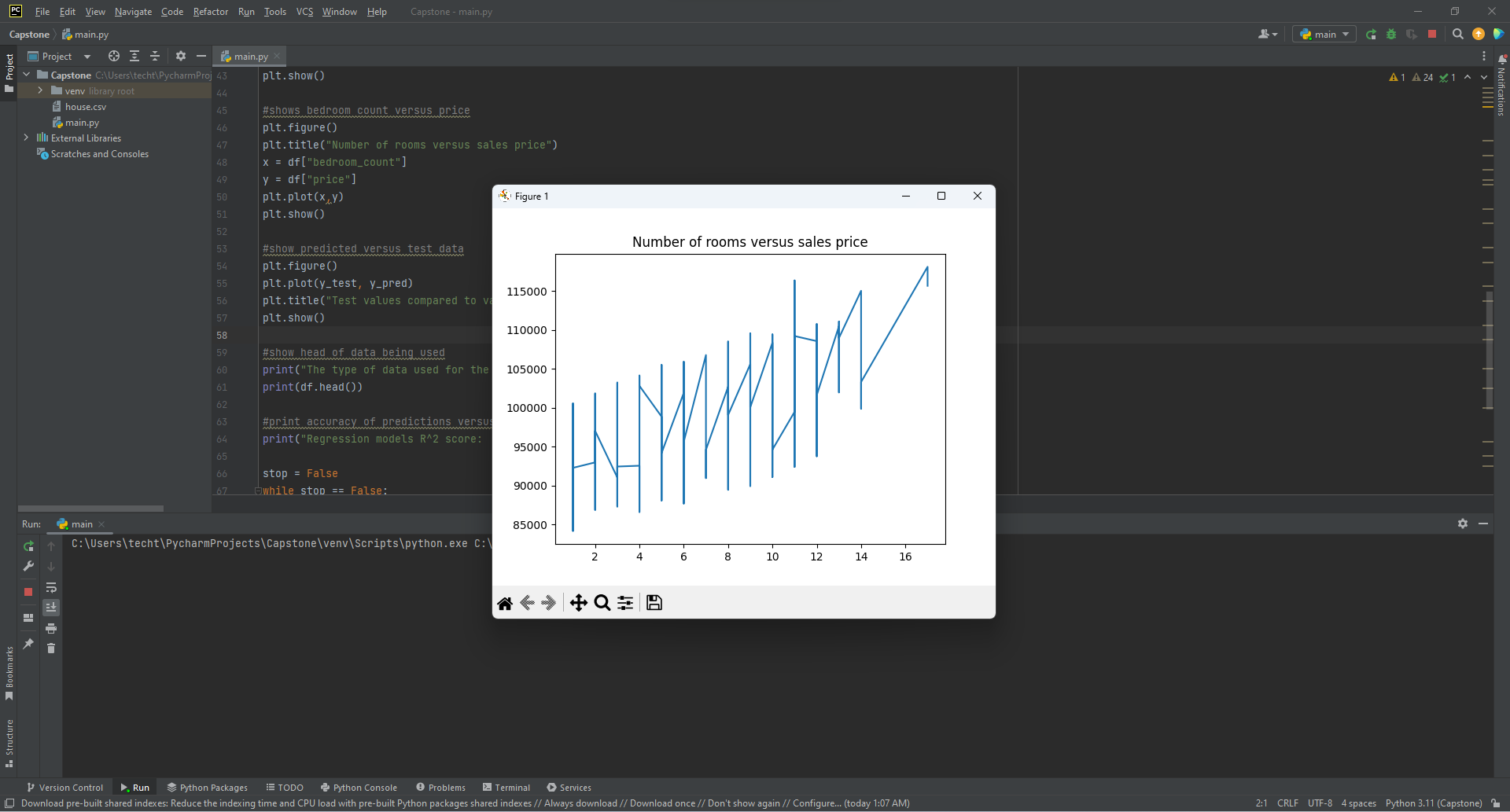
## Effective Visualization and Reporting

The model utilized to predict the sale price of a potential listing utilizes a variety of features concerning a property as inputs in the training and then prediction of property evaluation. The data used in this project contains information about key features of a property like the square meters of the property, how many rooms the property contains, how far the property is from the city center, what floor the property is on, etc. The descriptive representations provided for the user show a variety of figures to allow the user to visualize these key features in relation to selling price for a property. Each of the key features contained in the data are defining descriptors of a property and when evaluated in unison can provide an accurate evaluation of a property based on properties of similar characteristics. The following figures are presented to the user upon engaging the application.

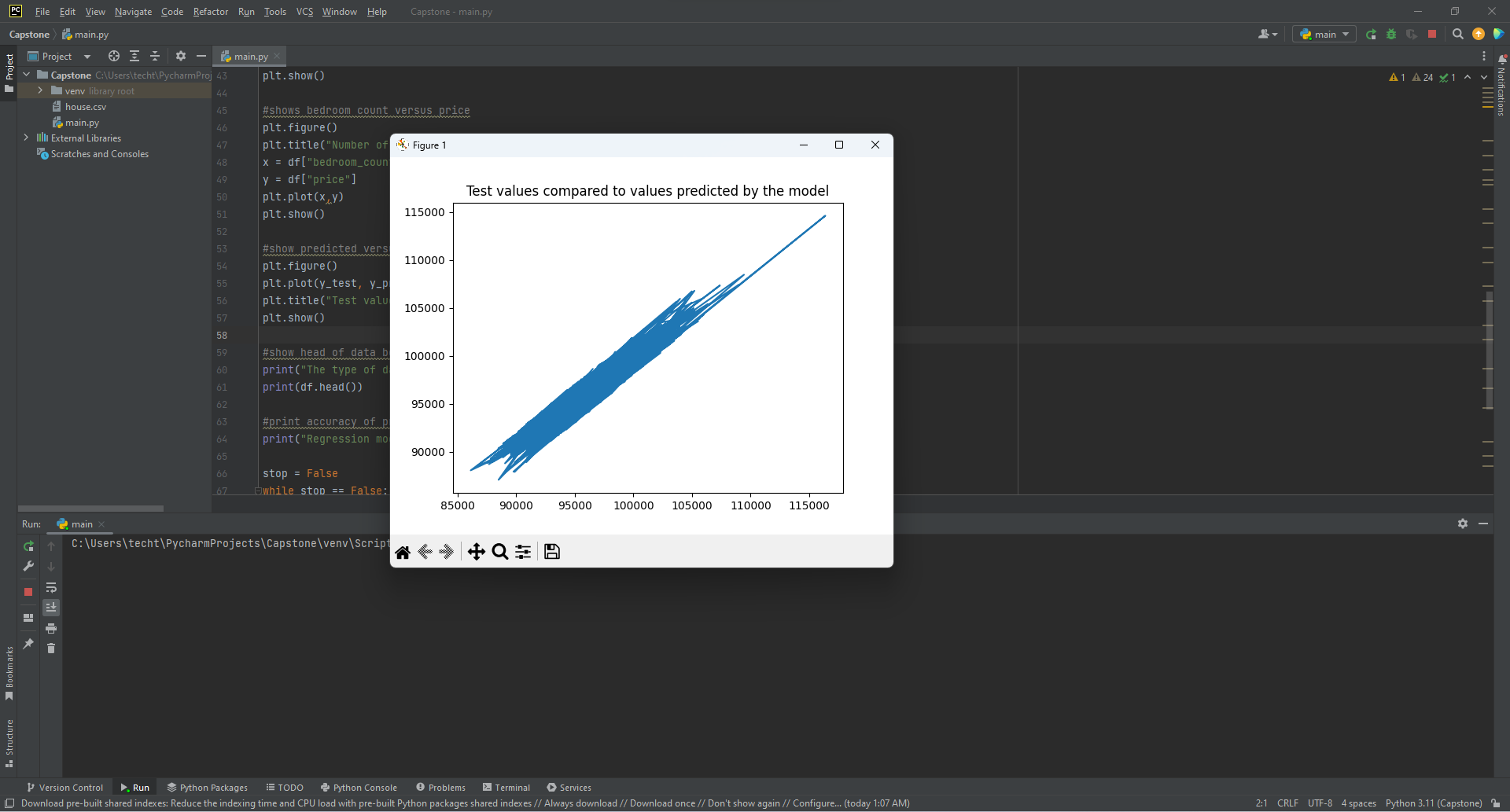
First there is a table showing the data in terms of the square meter of a property versus the sale price of the property:



The next provides a breakdown of the number of rooms associated with a property and its final sale price:



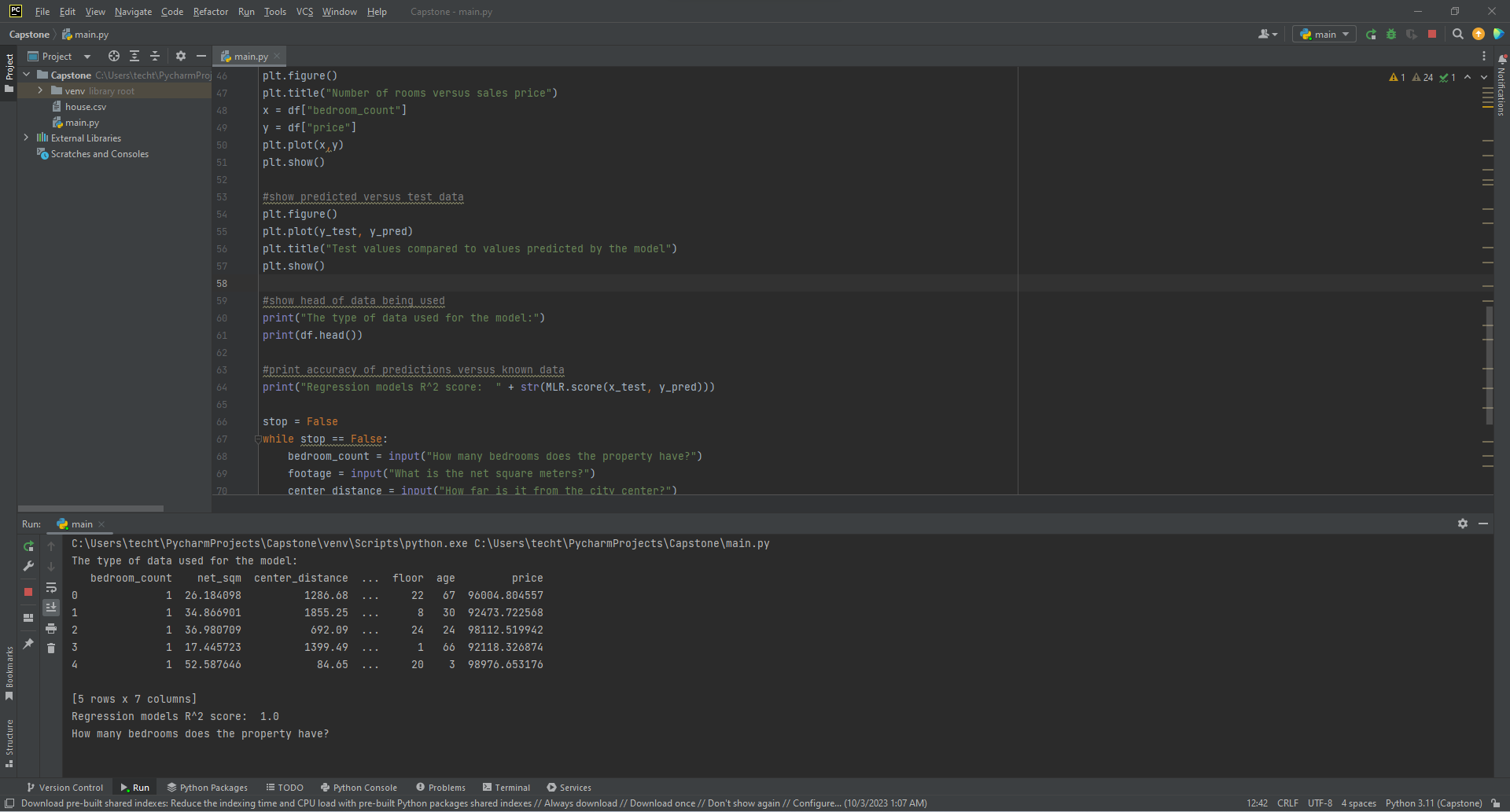
Finally, the program provides a visual analysis of the data being used to test the model. This visualization plots the predicted value from the model against the actual sale price contained within the test data site:



These tables show that there is a correlation between a variety of factors when a property is being evaluated to determine the price that it should be listed. Using this knowledge a Multiple Linear Regression model was utilized for the project. This model allows the model to incorporate each of these key features for a property in the model’s training in order for the model to generate the most accurate results when predicting the evaluation of a property based on various instances of the key features of a property utilized by the model.

## Accuracy Analysis

It was determined by the team that the best way to determine the accuracy of the model being utilized for the project was to calculate the R-Squared value of predicted values for a test set that had been sectioned off from the original data. This method compared the models predictions versus the actual sales record contained in the test data set and then generates a number between zero and one to determine accuracy, with one being the most accurate. The following results were shown during this analysis:



To further test the accuracy and determine that predicted values fell within the acceptable range of data being utilized by the model many tests were performed. The number of tests conducted makes it impossible to include all test results, however the following examples provide a greater understanding of this stage.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Application Testing

The application was tested utilizing a subset of the original data labeled as the test set. When training was completed using the training set, the test set was introduced to the model. The predictions that were generated by the model were then compared to the actual sale price contained within the test data set for each property. The R-Squared value based on this comparison was then calculated by the program and presented to the development team. This number ranges from zero to one. The closer the R-Squared number is to one the higher the accuracy of the model. The testing showed a high accuracy for the model being utilized, therefore no improvements were necessary.

## Application Files

The following items and libraries are required to execute the program:

* Python 3.11
  + Pandas libray
  + Matplotlib library
  + Scikit-learn library
* house.csv file

The submission provided a zip folder containing the over all project. This project zip file contains the house.csv file necessary for the program. For testing purposes the user would unzip the project folder into the PyCharms project folder.

## User Guide

1. Download PyCharm’s latest version.
2. Install Python’s latest version as the interpreter for PyCharm.
3. Install pandas using the terminal (Alt + F12) and the command pip install pandas.
4. Install matplotlib using the terminal (Alt + F12) and the command pip install matplotlib.
5. Add Scikit-learn library to the libray list by going to File >> Settings. Under Project:Python select the python interpreter. Click the + button and in the search bar type in scikit-learn. Select the library and then click install package.
6. Run program

## Summation of Learning Experience

* Describe how your previous experience (academic or professional) readied you for this project.
  + I am about to graduate with a Computer Science Bachelor’s degree. This provides me with the technical skills to perform not only the coding for the project but its underlying components.
* Describe any additional learning or resources needed to complete this project.
  + There were no additional learning or resources needed to complete this project.
* Describe how this experience contributed to your concept of lifelong learning.
  + Based upon this project you need to be redundant in the career field. This project asked to described the problem of the project in each section of the project minus part C which is thte application. Various sections of the project require the participant to answer the same questions so redundancy is apparently an important part of the career. \*SIGH\*