**Title:** Exploratory Data Analysis in Python via Pandas and Matplotlib

**Author:** Johnny L. Hopkins / hopkins28160

**Track Category:** Data Science

**Subcategory:** Pandas, Matplotlib

**Tags:** Python, Pandas, Matplotlib, Jupyter, Anaconda, data, analysis, Ames, housing, dataset

**Read Time:** 15 minutes

**Banner Image:** <https://www.pexels.com/photo/snow-covered-trees-in-winter-310386/>

**Content:**

Over the past few decades, many researchers and scientists have generated huge amounts of data. One of the earlier undertakings was the Human Genome Project, where molecular biologists succeeded in mapping the entire human genome. As a result, the discipline of bioinformatics was born. People give ratings for products purchased on Amazon and give reviews of businesses on Yelp. Business generate financial data used for end of fiscal year reports and decision-making. How does a data scientist analyze this data and make sense of it all? To achieve this task, a data scientist goes back to his or her statistics courses and perform exploratory data analysis. This analysis takes data sets and looks at general patterns. We are not concerned about decision making or model building at this point, the goal is to describe what we have in our data sets. Fortunately, we have tools such as R and Python plus non-specialist software such as Orange (based on Python) and KNIME. In this post, I will use Python to perform exploratory analysis on the Ames Housing dataset. The tools I will use include Panadas, Matplotlib, and Jupyter. The code examples will be presented as a Jupyter notebook which will be available on my GitHub repository.

The Ames Housing dataset was compiled by Dean Cook and was designed for use in data science education. This dataset list 79 explanatory variables describing all aspects of homes in Ames, Iowa and is use for predicting the selling prices for these homes using regression techniques. This post uses the Ames Housing dataset to perform exploratory data analysis including plots and descriptive statistics.

I am using both Pandas and Matplotlib for the analysis. Pandas is a library used for data analysis and is frequently utilized in data science and data analytics. Matplotlib is Python’s library for creating and generating scientific graphs. Together with Jupyter, a notebook-bases interface for scientific computing, I will do a simple analysis on the Ames Housing data. A key function in Pandas is the describe()function, which list the important descriptive statistics in the data set. When using Pandas, data sets are exported into DataFrame(s) to facilitate data analysis. Then, Matplotlib is used to generate plots with the most common being histograms. Since this data set is used for predicting prices, I could generate a scatterplot, but this is beyond the scope of this article.

To understand Exploratory Data Analysis, you must understand the following objectives:

* To suggest possible hypotheses about our data
* Form the basis of statistical inference by assessing assumptions about our data
* Determine the statistical tools and techniques used in our analysis of datasets
* Through surveys, experiments, simulations, etc., give a basis for further data collection

Exploratory Data Analysis is a method used to develop statistical thinking and to think critically about assumptions made from looking at a dataset. Graphical techniques include the following:

* Box plots
* Histograms
* Scatterplots
* Pareto charts
* Stem-and-leaf plots
* Etc.

The plots mentioned are introduced in the first few weeks of an elementary statistics course or business statistics course. Numerical techniques for Exploratory Data Analysis include calculating the:

* Mean
* Median
* Mode
* Quartiles
* Variance and Standard Deviation
* Interquartile Range (IQR)
* Maximum and minimum values of a dataset

Again, these techniques are encountered in elementary statistics. The first half of an elementary statistics course is an exercise in Exploratory Data Analysis since we want to describe the characteristics of a dataset before making inferences on our data and building models based on our datasets.

Having explained the premises behind Exploratory Data Analysis, I will now turn our attention to the Ames Housing dataset.

Now, we are ready to explore our data set. The notebook for this post can be found at this [link](http://github.com/jlhopkins28160/TopcoderTHRIVE/Ames.ipynb). For starters, we must import the data into our code using read\_csv(). Here are our imports:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

%matplotlib inline

We must import the NumPy library because the Ames Housing dataset is read into a data frame. The next important declaration is %matplotlib inline, which allows us to output our graphs in Jupyter. Next, we use the read\_csv() function to read in our data. After we read in the data, we output its contents using the print() function, which gives us this output:

Order PID MS SubClass MS Zoning Lot Frontage Lot Area Street \

0 1 526301100 20 RL 141.0 31770 Pave

1 2 526350040 20 RH 80.0 11622 Pave

2 3 526351010 20 RL 81.0 14267 Pave

3 4 526353030 20 RL 93.0 11160 Pave

4 5 527105010 60 RL 74.0 13830 Pave

... ... ... ... ... ... ... ...

2925 2926 923275080 80 RL 37.0 7937 Pave

2926 2927 923276100 20 RL NaN 8885 Pave

2927 2928 923400125 85 RL 62.0 10441 Pave

2928 2929 924100070 20 RL 77.0 10010 Pave

2929 2930 924151050 60 RL 74.0 9627 Pave

Alley Lot Shape Land Contour ... Pool Area Pool QC Fence Misc Feature \

0 NaN IR1 Lvl ... 0 NaN NaN NaN

1 NaN Reg Lvl ... 0 NaN MnPrv NaN

2 NaN IR1 Lvl ... 0 NaN NaN Gar2

3 NaN Reg Lvl ... 0 NaN NaN NaN

4 NaN IR1 Lvl ... 0 NaN MnPrv NaN

... ... ... ... ... ... ... ... ...

2925 NaN IR1 Lvl ... 0 NaN GdPrv NaN

2926 NaN IR1 Low ... 0 NaN MnPrv NaN

2927 NaN Reg Lvl ... 0 NaN MnPrv Shed

2928 NaN Reg Lvl ... 0 NaN NaN NaN

2929 NaN Reg Lvl ... 0 NaN NaN NaN

Misc Val Mo Sold Yr Sold Sale Type Sale Condition SalePrice

0 0 5 2010 WD Normal 215000

1 0 6 2010 WD Normal 105000

2 12500 6 2010 WD Normal 172000

3 0 4 2010 WD Normal 244000

4 0 3 2010 WD Normal 189900

... ... ... ... ... ... ...

2925 0 3 2006 WD Normal 142500

2926 0 6 2006 WD Normal 131000

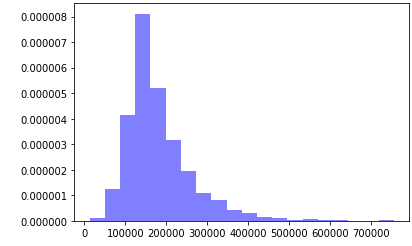
2927 700 7 2006 WD Normal 132000

2928 0 4 2006 WD Normal 170000

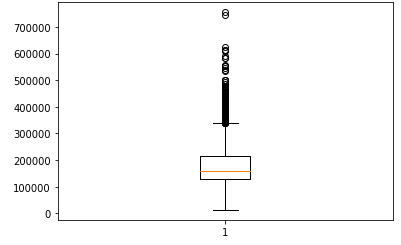
2929 0 11 2006 WD Normal 188000

[2930 rows x 82 columns]

To get rid of the NaN variables and do summary statistics, we call the describe() function which outputs the following stats: count, mean, std (standard deviation), min, 25%, 50%, 75% percentiles, and max. Next, I will generate a histogram for the SalePrice variable as it is the variable of interest in modeling sales prices (I will not generate the predictive models for the Ames dataset as the purpose here is to explore our data and make general assumptions). I use the hist() function in matplotlib to plot the SalePrice data and look at variations in pricing:



In looking at the histogram, we have a left-skewed distribution of prices, meaning that most homes sell between $50,000 and $400,000 roughly. Our mean of SalePrice was $180,796 (rounded to the nearest dollar) and the standard deviation was $79,886.69 (rounded to two decimal places. The boxplot shows the variation in prices for houses in Ames, Iowa:



Looking at the boxplot, the mean is denoted in red and the circles represent our outliers in our data set. The handlebars indicate that our prices range from $0 to $400,000. Our histogram shows rough the same range of prices. If a distribution is right-skewed, the mean > median of our dataset. The median for SalePrice is $160,000, which is calculated using the median() function in Pandas.

We have used Exploratory Data Analysis on the Ames Housing dataset. We focused on one variable, SalePrice, and generated basic statistics and plots for our housing prices. Through our analysis, the median sales price of a house in Ames, Iowa is $160,000 and the mean sales price is $180,796. Because the mean sales price > median sales price, the SalePrice data is skewed left. This shows that our prices range between roughly $50,000 and $400,000, which skews our data. This is evident in both our histogram and boxplot. This is how you do Exploratory Data Analysis in Python. You are encouraged to download the [Jupyter notebook](http://github.com/jlhopkins28160/TopcoderTHRIVE/Ames.ipynb) for the article and explore other variables in the Ames Housing dataset. You can also use the notebook to do exploratory analysis on other datasets of interest.