Bella Davies

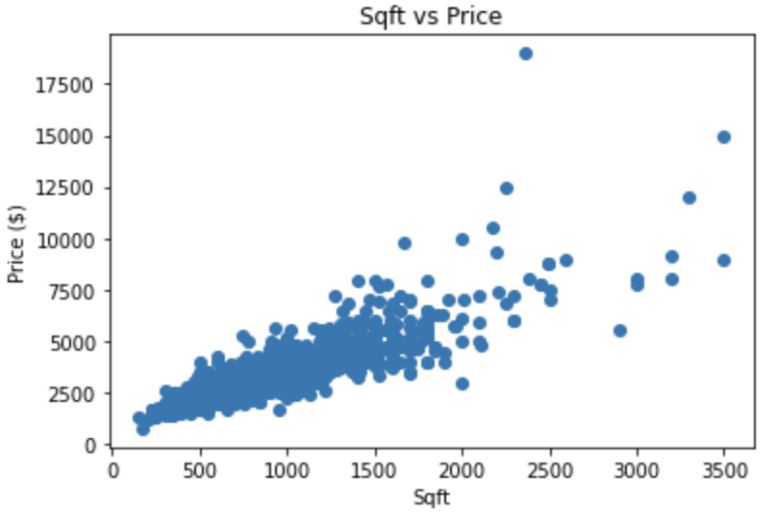
DS 210: Programming for Data Science

Professor Kontothanassis

15 December 2022

DS 210 Final Project Report

In this final project, my goal is to analyze the dataset imported from Kaggle titled, "Housing Prices in San Francisco (Craigslist)". The cleaned dataset from the zipfile downloaded from Kaggle, sf_clean.csv, contains the data for housing prices from Craigslist in the city of San Francisco. Within the cleaned dataset, each row describes the price, square footage, number of bedrooms, number of bathrooms, whether there is laundry in the unit, whether dogs, cats, both, or none are allowed, the housing type, the type of parking, and the hood district. For the purposes of this project, I will be looking at the first two columns, "price" and "sqft". By assigning the first two columns, "price" and "sqft" to a data structure called DataFrame, I am able to use plotly to create a scatter plot of the data points with the "sqft" column as the x values, and "price" as the y values. Upon looking at the visualization for the plot of sqft by price, we can see that the scatter plot has a general upwards trend.



The algorithm used in this project is linear regression. The scatter plot of the data for Sqft vs Price shows a general upward trend, and by implementing linear regression using polyfit_rs, we can find the general trendline for the dataset using it's slope and intercept. The result of my implementation of linear regression using polyfit_rs on the cleaned data set "sf_clean.csv" describing housing prices in San Francisco shows that the trendline of this data has a slope of 2.723 and a y-intercept of 935.369.

In conclusion, the implementation of linear regression tells us that for each square footage increase in housing, the housing price will increase by $2.72.